Fra speilmetoden til automatisk ekstrahering av et betydningstagget korpus for WSD-formål

Gunn Inger Lyse

juni 2003

Hovedoppgave i datalingvistikk
Seksjon for lingvistiske fag
Universitetet i Bergen

Innhold:

Abstract

This thesis adresses the lack of sense-annotated corpora as a background resource for Word Sense Disambiguation (WSD). The most promising approach to WSD is generally considered to be corpus-based, supervised machine learning methods. In this approach, a sense-tagged training corpora provides example instances which illustrate the relation between a given word sense and its typical context. However, supervised learning has proven to be limited as a larger-scale alternative, because sense-tagged corpora need to be manually tagged, which is costly and time-consuming. Consequently, it is desirable to investigate methods to overcome this knowledge acquisition bottleneck.

This thesis suggests a method which automatically extracts a finite, sense-tagged corpus. Although the method is only tested on one ambigous lemma within this thesis, the method is in principle expected to be applicable for extracting sense-tagged corpora for all ambigous words within the vocabulary of a given language. The presented method is based on translational correspondences in a parallel corpus, sorted by meaning by a "semantic mirroring" method (Dyvik, 1998/2002). The chief goal of the thesis is to explore the presented method's potential as an alternative to a manual sense-tagging of corpora. The results are first evaluated manually. Then follows a practical evaluation, by applying the automatically sense-tagged corpus as training material for a supervised learning algorithm. The results reveal that the presented approach methodically seems promising, indicating a good potential for further exploration.

Sammendrag

Utgangspunktet for denne oppgaven er mangelen på tilgjengelige betydningstaggede korpora som bakenforliggende ressurs for automatisk orddisambiguering (Word Sense Disambiguation; WSD). WSD-tilnærmingen som per i dag regnes som mest lovende, korpusbasert overvåket maskinlæring, har vist seg begrenset i praktisk bruk fordi den forutsetter tilgang på et betydningstagget treningskorpus som eksemplifiserer sammenhengen mellom en ordbetydning og dens typiske kontekst. Betydningstaggingen av slike treningskorpora må i dag utføres manuelt, hvilket er kostbart og tidkrevende arbeid. Det er derfor ønskelig å undersøke metoder for å automatisere dette arbeidet.

Denne oppgaven foreslår en metode som ekstraherer et finitt, betydningstagget korpus automatisk. Selv om metoden av tidshensyn kun er testet på ett flertydig norsk lemma innenfor rammene av denne oppgaven, er metoden prinsipielt forventet å kunne ekstrahere betydningstaggede korpus for alle flertydige ord innenfor et språks vokabular. Metoden er basert på oversettelseskorrespondanser i et parallellkorpus, som er sortert etter betydning ved speilmetoden (Dyvik, 1998/2002). Målet er å undersøke den presenterte metodens potensial som alternativ til en manuell betydningstagging av et korpus. Evalueringen foregår først manuelt. Dernest følger en praktisk evaluering, ved å anvende metodens betydningstaggede korpus som treningsmateriale i en overvåket maskinlæringsalgoritme. Resultatene indikerer at oppgavens presenterte tilnærming metodisk sett synes lovende, og at metoden derfor har et stort potensial for videreutvikling.

Forord

Denne hovedoppgaven inngår i forskningsprosjektet "Fra parallellkorpus til ordnett" ved at oppgaven anvender metoden "semantiske speil" som grunnlag for eksperimenter med automatisk orddisambiguering. Forskningsprosjektet er tilknyttet Universitetet i Bergen og HIT-senteret, og er finansiert av Meltzerfondet og Norges forskningsråd, som også har tildelt et studentstipendium for denne oppgaven.

Jeg vil rette en stor takk til prosjektleder professor Helge Dyvik, som bidro med selve ideen til denne oppgavens presenterte metode for automatisk betydningstagging, og som har vært hovedveileder for denne oppgaven. Hans praktiske hjelp og innsiktsgivende innspill underveis har vært uvurderlige for utviklingen av oppgaven. Likeledes vil jeg takke de andre involverte i ordnett-prosjektet. Spesielt takkes Sindre Sørensen for tålmodig bistand under LISP-implementeringen av oppgavens eksperimenter, og Martha Thunes som har bidratt sterkt som diskusjonspartner med hensyn til oppgavens anvendte eksperingsprinsipper. Videre rettes en stor takk til min biveileder, professor Koenraad de Smedt, som har bistått med praktisk hjelp og verdifulle innspill i forbindelse med maskinlæringsdelen av oppgaven.

1. Introduksjon

1.1 Innledning

Word Sense Disambiguation (fra nå av: WSD) går ut på å fastslå hvilken betydning av et flertydig ord som er brukt i en spesifikk kontekst, og refereres ofte til som et av de store uløste problemer innenfor forskningsfeltet språkprosessering (fra nå av: NLP[1]) i dag (Resnik & Yarowsky, 1997/2000; Ide & Véronis, 1998; Stevenson, 2003).

Som Stevenson (2003) påpeker, har forskning rundt WSD fra 50-tallet og til i dag møtt på ulike former av hva Gale et al. (1992) betegner som "the knowledge acquisition bottleneck": Problemet med å finne tilgjengelige og anvendbare kunnskapsressurser for å utføre automatisert orddisambiguering. Følgelig kan forskning rundt tilnærminger til WSD grunnleggende sett sies å bestå i å utforske hvilke tilgjengelige ressurser som er egnet som bakenforliggende ressurs for WSD.

"The knowledge acquisition bottleneck" nevnes i dag først og fremst i forbindelse med eksperimenter rundt såkalte overvåkede maskinlæringsmetoder for WSD, som er den tilnærmingen man resultatmessig regner som det mest lovende paradigmet for WSD (Resnik & Yarowsky, 1997/2000, Agirre & Martinez 2001, Diab & Resnik 2002, Escudero et al. 2000). Overvåket WSD forutsetter et treningskorpus hvor hver betydning av ordet/ordene som skal disambigueres på forhånd er korrekt tagget for betydning i kontekst. På grunnlag av dette korpuset "lærer" systemet sammenhengen mellom en viss ordbetydning og dens typiske kontekst, og kan overføre denne (generaliserte) lærdommen for å disambiguere forekomster av ordet i nye kontekster.

Den semantiske taggingen av et slikt treningskorpus må per i dag utføres for hånd eller semiautomatisk, noe som er kostbart og tidkrevende arbeid, og som dermed begrenser størrelsen på korpus mer enn hva som er ønskelig for en god maskinlæringseffekt. I tillegg kommer at den virkelige arbeidsmengden ville bestå i å skaffe betydningstaggede treningskorpora for alle flertydige ord i vokabularet, for å kunne gjennomføre overvåket WSD i større skala. Det er derfor ønskelig å undersøke mulighetene for å utvikle en metode som automatiserer betydningstaggingen av et korpus.

Denne oppgaven presenterer en metode for automatisk ekstrahering av et slikt betydningstagget treningskorpus. Metoden anvender oversettelseskorrespondanser som på

[1] NLP er en forkortelse for Natural Language Processing.

forhånd er sortert etter betydning ved speilmetoden, en metode utviklet av Dyvik (1998/2002) innenfor forskningsprosjektet "Fra parallellkorpus til ordnett". Dette betydningstaggede korpuset vil så gis som treningsmateriale til en maskinlæringsalgoritme for overvåket WSD.

Før vi konkretiserer hypotesen og metoden i denne oppgaven, er det formålstjenlig å først gi en bakgrunnsforståelse av denne oppgavens konkrete problemstiling i et forskningsperspektiv. (1.2) gir en nærmere introduksjon av problemet WSD (1.2), fulgt av en diskusjon av WSDs relevans i et større NLP-perspektiv (1.3). I (1.4) følger en kort oversikt over hovedtilnærminger til WSD, med fokus på såkalte "hybrider" som retter seg mot mangelen på betydningstaggede korpora. (1.5) presenterer denne oppgavens eksperiment.

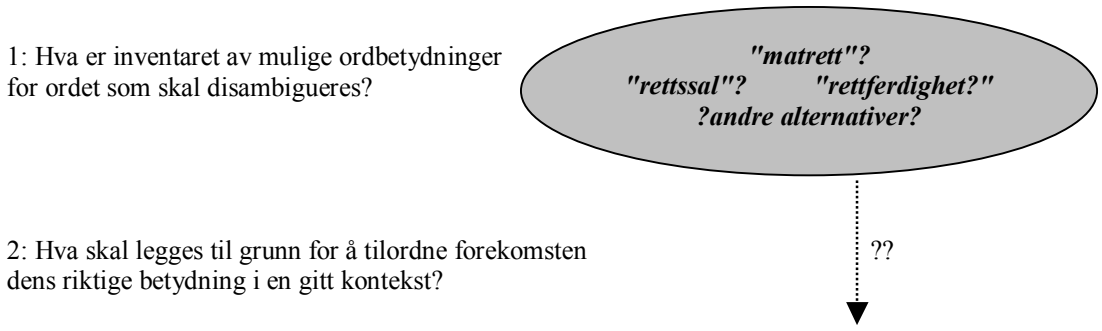
1.2 WSD som problem

Beskrivelsen av WSD som går igjen i litteraturen er som i introduksjonen i (1.1) ovenfor: WSD er prosessen med å automatisk identifisere hvilken betydning et flertydig ord har i en viss kontekst (Stevenson, 2003). Men hva *innebærer* dette? Forutsetningene for å utføre automatisk orddisambiguering kan sies å involvere to ulike aspekter.

For det første forutsetter automatisk orddisambiguering en kilde til informasjon om ords *potensielle* betydninger: Skal et system kunne disambiguere et gitt ord, så må systemet få (eller kunne utarbeide automatisk) kunnskap om hvilket inventar av ordbetydninger det har å velge mellom. For det andre innebærer enhver tilnærming til orddisambiguering å se på konteksten som det aktuelle flertydige ordet inngår i. Da er det på den ene siden et spørsmål om hvilken kunnskapskilde som er anvendbar for å avgjøre betydning, altså hvilken kilde som kan tjene som systemets referansepunkt for sammenligning. På den andre siden kommer et empirisk spørsmål om hva som er å regne som relevant kontekst: Skal man benytte såkalt lokal kontekst, det vil si den umiddelbare konteksten (kollokasjoner) rundt ordet; eller nøkkelord (antatt karakteristiske ord) i en større kontekst, eller en kombinasjon av disse?

Kunnskapsaspektene ved WSD er illustrert i figur 1.1 under.

Figur 1.1: De to kunnskapsaspektene ved Word Sense Disambiguation (WSD)



*På kvelden ble det servert en tradisjonell norsk **rett**.*

Som figuren illustrerer, er det altså for det første snakk om en kunnskapskilde for å vite noe om selve ordbetydningene, og for det andre kreves en ressurs og en tilnærming for å utføre selve disambigueringen av et flertydig ord i en viss kontekst. Felles for begge aspekter er at det er et spørsmål om å finne frem til tilgjengelige og anvendbare bakenforliggende kunnskapskilder for WSD.

1.3 WSD som mål i et større forskningsperspektiv

Interessen for WSD går helt tilbake til utviklingen av maskinoversettelsessystemer på 50-tallet: Hvordan skal maskinen velge den riktige oversettelsen av et ord i en gitt setning, hvis ordet er oppført med flere mulige oversettelser? Som eksempel kan man tenke seg et maskinoversettelsessystem som skal oversette den norske setningen (1) under til engelsk (2). Uten et system for disambiguering kan ikke oversettelsessystemet slå fast om det flertydige norske substantivet *rett* her burde oversettes med *dish* eller et "rettslig" relatert ord som *right*. (Gal oversettelse markert med * foran.)

(1) *På kvelden ble det servert en tradisjonell norsk **rett**.*
(2) *In the evening a traditional, Norwegian **dish**/***right** was served.*

Ords potensielle flertydighet utgjør et problem innenfor de fleste NLP-systemer i større eller mindre grad. I tillegg til maskinoversettelsessystemer kan vi for eksempel nevne systemer for informasjonssøking/-henting, tekstsammendrag, syntaktisk parsing og for talesyntese/-gjenkjenning.

Wilks & Stevenson (1996) påpeker derfor at WSD "ikke er et mål i seg selv": Det overordnede målet for forskning rundt WSD er å komme frem til tilnærminger til WSD som kan være anvendbare innenfor "større" NLP-systemer, som systemer for maskinoversettelse. Vi kan således formulere det som at WSD-tilnærminger prinsipielt sikter mot å være metoder for å kunne disambiguere en potensielt infinitt mengde av nye forekomster av et gitt flertydig ord – Med andre ord slik problemstillingen vil foreligge innenfor "overordnede" NLP-systemer, hvor det ikke på forhånd er gitt hvilke flertydige ord som må løses opp og hvilke kontekster de forekommer i.

Wilks (2000) reflekterer over hvorvidt automatisk orddisambiguering overhodet er et oppnåelig mål, og i hvilken grad selv et perfekt WSD-system vil ha en virkelig nytteverdi i et NLP-system. Det første spørsmålet har å gjøre med kompleksiteten som ligger i ordsemantikkens natur. Som Wilks påpeker, skaper mennesker stadig nye mulige kontekster for et ord, noe som i sin tur skaper stadig nye betydningsnyanser. Det er sjelden et problem for mennesker å finne den riktige tolkningen av et ord selv i en splitter ny kontekst, men å definere en kontekstuavhengig grense mellom flere betydninger av et ord er en oppgave som selv ikke trenede leksikografer alltid evner å løse konsekvent. Man kan derfor spørre seg om orddisambiguering overhodet er å regne som en oppgave som *kan* automatiseres.

Wilks (2000) argumenterer i den forbindelse for at den menneskelige evnen til å produsere og forstå ord i helt nye ordbetydninger skiller WSD-oppgaven fra en språkprosesseringsoppgave som f.eks. part-of-speech-tagging (fra nå av: pos-tagging). Med dette mener han at et leksems inndeling i ordbetydninger sjelden er like veldefinerbar og avgrenset som det samme leksemets riktige syntaktiske kategori (pos). Definisjonen av betydningsdistinksjoner for et leksem varierer som regel fra leksikon til leksikon og fra språkbruker til språkbruker, og i tillegg kan en definisjon vanskelig ta høyde for menneskelig språklig kreativitet. Wilks påpeker at dette legger en begrensning på mulighetene for evaluering og på effektiviteten man kan forvente av et WSD-system, og han mener derfor at WSD må anses som langt mer enn "nok en språkprosesseringsoppgave som skal løses". Gitt at de beste systemene for WSD per i dag har en nøyaktighet på over 90% (Yarowsky, 1995; Stevenson & Wilks, 2001) er det ingen grunn til å mene at WSD ikke er mulig, men Wilks har nok rett at i at man heller ikke skal ha for høye forventninger.

Det andre spørsmålet Wilks (2000) tar opp er i hvilken grad WSD-system har noen virkelig nytteverdi som en integrert del av større systemer for språkprosessering. Resnik&Yarowsky (1997/2000) observerer at potensialet for WSD varierer etter hvilken type

NLP-system det er ment å integreres i. For eksempel påpeker de at et system for talegjenkjenning vil ha større nytte av tilgang på klasser av kontekst snarere enn ordbetydningsklasser. Videre påpeker de muligheten for at selv perfekt orddisambiguering kan vise seg å ha begrenset verdi i praksis innen informasjonshenting (IR; Information Retrieval), fordi en søkestreng med flere nøkkelord kan være nok i seg selv til å disambiguere (gitt at det er liten sannsynlighet for at alle nøkkelordene er flertydige). Resnik & Yarowsky (1997/2000) peker ut maskinoversettelse som den NLP-oppgaven hvor WSD har et størst brukspotensial. Allerede i 1959 slo filosofen Bar-Hillel (1959) fast at leksikalsk flertydighet utgjør den store barrieren med hensyn til maskinoversettelsessystemets kvalitet. Imidlertid er det også for maskinoversettelsessystemer verdt å huske på at man ikke bør ha for høye forventninger til graden av forbedring selv med et perfekt system for WSD. Hvis et system for maskinoversettelse oversetter mellom to språk med en korrekthet på la oss si 94 %, så er det ikke nødvendigvis slik at de ”manglende” 6 % alle bunnet i problemer med tvetydighet. For eksempel kunne disse 6 prosentene også bunne i gal setningsanalyse innenfor målspråket eller kildespråket, eller i mangler i leksikon som ikke nødvendigvis gikk på flertydighet.

Wilks (2000) påpeker imidlertid at den eneste måten for å lære mer om WSDs sanne potensial for å forbedre NLP-systemer, er å holde fortsette forskningen rundt WSD slik at vi empirisk kan teste WSD-metoders anvendelighet i NLP-systemer.

Neste delkapittel gir en kort introduksjon til hovedtilnærmingene til WSD.

1.4 En oversikt over tilnærminger til WSD

Dette delkapittelet presenterer de tre hovedtilnærmingene til WSD som man regner med i dag med hensyn til bruk av maskinlesbare bakenforliggende ressurser (Stevenson, 2003). De to "etablerte" tilnærmingene man regner med vil vies en generell presentasjon i (1.4.1) og (1.4.2). Som vi vil se, har disse to tilnærmingene fordeler som utfyller hverandre, og dette har ledet til en "hybrid"-kategori (1.4.3) som kombinerer de to hovedtilnærmingene man tidligere har regnet med.

Formålet med kapittelet er å plassere den spesifikke problemstillingen med manglende betydningstaggede korpora innenfor WSD som forskningsfelt, og å vise hvordan "hybrid"-eksperimenter kan karakteriseres som den mest aktuelle tilnærmingen for å løse dette problemet. Innenfor rammene av denne oppgaven vil derfor fokus legges på "hybrider", og denne tilnærmingen vil konkretiseres ved noen utvalgte tidligere forskningsforsøk. For en mer utførlig oversikt over forskning innenfor WSD refereres leseren til Ide & Véronis (1998).

1.4.1 Kunnskapsbaserte tilnærminger til WSD

Før maskinlesbare kunnskapsressurser som ordbøker, tesauri og leksika ble elektronisk tilgjengelig på 80-tallet, måtte forskning rundt WSD stort sett basere seg på begrensende manuelt oppbygde ressurser som f.eks. seleksjonsrestriksjoner, som sjelden dekket et større vokabular. Dette la åpenbart store begrensninger på utstrekningen av eksperimenter som var praktisk gjennomførbare, og man kan således si at forskningen rundt WSD tidlig erfarte flaskehalsen som mangelen på tilgjengelige ressurser medførte. Ide & Véronis (1998) karakteriserer det derfor som et vendepunkt for forskning rundt WSD da mer omfattende maskinlesbare leksikalske ressurser og korpora ble elektronisk tilgjengelige på henholdsvis 80- og 90-tallet.

Tilgangen på slike leksikalske ressurser åpnet for eksperimenter rundt WSD som med en samleterm kalles kunnskapsbaserte tilnærminger. Karakteristisk for slike tilnærminger er at de anvender "eksplisitte" leksikalske kunnskapsressurser som informasjonskilde for WSD. Den anvendte leksikalske ressursen supplerer "eksplisitt" predefinerte betydningsskiller som anvendes i orddisambigueringen, og ressursen tjener i tillegg som bakenforliggende kilde for selve orddisambigueringen. Under følger en oppsummering av de tre hovedtypene leksikalske kunnskapsressurser.

1.4.1.1 Maskinlesbare ordbøker

En maskinlesbar ordbok (en MRD, som står for *Machine-Readable Dictionary*), er organisert slik at et hovedoppslag består av et lemma som inndeles i flere underbetydninger, og hver underbetydning er tilknyttet en betydningsdefinisjon. Orddisambiguering foregår da med utgangspunkt i ordene som befinner seg i betydningsdefinisjonene. Den essensielle grunnideen er som følger: Gitt et flertydig ord *x*, som ifølge en MRD er oppført med *n* underbetydninger med hver sine tilhørende definisjoner. For å disambiguere en forekomst av *x*, blir ordene i *x*s betydningsdefinisjoner sammenlignet med ordboksdefinisjonene til hvert av ordene i konteksten til den aktuelle ordforekomsten.

Ide & Véronis (1998) legger fram følgende fordeler og ulemper ved bruken av MRD for WSD: Fordelen ved ordbøker er at de gir tilgang til et stort sett flertydige ord (med ferdig definerte betydningsskiller), noe som potensielt åpner for orddisambiguering utført på alle ord i en tekst som der finnes MRD-oppslag for. Imidlertid utgjør i praksis en kort ordboksdefinisjon som regel et alt for spinkelt grunnlag for å tilegne seg tilstrekkelig kunnskap om de mulige kontekster som en gitt betydning kan assosieres med. Videre påpeker Ide & Véronis at ordbøkers manuelt definerte betydningsinndelinger ofte er inkonsekvente. I tillegg kan det være et praktisk problem at informasjonen i en MRD er kodet med tanke på at mennesker (og ikke maskiner) skal bruke dem.

1.4.1.2 Tesauri

Tesauri tilbyr på sin side et nettverk av semantiske assosiasjoner mellom ord (primært mht. synonymi). Enkelte tesauri definerer også sett av semantiske kategorier for ordbetydningene (f.eks. HUMAN eller OBJECT) som potensielt er av verdi for språkprosessering.

Likevel anses ikke tesauri som en svært lovende tilnærming til WSD. Hovedgrunnen til dette er ifølge Ide & Véronis (1998) at tesauri, som MRDer, primært er ment som en ressurs for mennesker. Imidlertid er det usikkert om de er tilstrekkelig konsekvente til å utgjøre en optimal kilde til informasjon om ordrelasjoner for maskinelle formål.

1.4.1.3 Leksika

Den mest kjente og brukte leksikalske ressursen for WSD er WordNet (Miller, 1998). Dette manuelt bygde leksikonet for engelsk inneholder definisjoner på alle ords gitte betydninger (som i en ordbok), og for substantiver defineres i tillegg hierarkisk organiserte sett av synonymer (kalt *synsets*). Et *synset* består av nærrelaterte ord som alle kan knyttes til et gitt leksikalsk konsept (som i et tesaurus). Disse konseptene lenkes sammen ved semantiske relasjoner som hyperomymi/hyponymi, antonymi og meronymi, og underbegreper arver trekk fra overbegrepene (*lexical inheritance system*).

Den semantiske informasjonen i WordNet kan f.eks. anvendes for å utføre orddisambiguering basert på hvor nært relatert to ord er. Resnik (1995) har brukt WordNet til

å beregne et mål på semantisk likhet for substantiver ved å kalkulere to ords "shared information content". Denne WSD-tilnærmingen innebærer å finne det hierarkisk laveste konseptet som subsumerer (er et overbegrep til) begge ord: Jo mer spesifikt konseptet som subsumerer begge ord er, jo nærmere er de to ordene relatert semantisk. På denne måten kan ord i konteksten til det flertydige ordet *x* måles mot hvilken betydning av *x* som er nærmest.

Kritikken mot WordNet er at dette ordnettet anvender svært fingrenede betydningsdistinksjoner, noe som kompliserer en WSD-algoritmes jobb med å finne ut hva som skiller én betydning fra en annen. Semantiske ordnett anses likevel generelt som en verdifull ressurs for orddisambiguering. Som vi skal se i (1.4.3) og i denne oppgavens presenterte metode, har de f.eks. et potensial som bakenforliggende ressurs for å betydningstagge korpora som treningsmateriale for korpusbasert WSD.

1.4.1.4 Oppsummering

Kunnskapsdrevne tilnærminger til WSD har generelt den fordelen at de supplerer et inventar av potensielle betydninger for et flertydig ord. En innvending mot mange leksikalske ressurser sine betydningsskiller er likevel at de har en tendens til å være for fingrenede, blant annet som en følge av maskinlesbare leksikalske ressurser primært er rettet mot menneskelig forståelse av bruken av et ord. Med hensyn til å bruke slike kunnskapsressurser som kilde for selve orddisambigueringen, viser det seg ofte at kilden ikke har tilstrekkelig informasjon som er anvendbar for å avgjøre et ords betydning i en gitt kontekst. Vi vil derfor se i (1.4.3) at leksikalske kunnskapsressurser i dag gjerne anvendes kun for å supplere forhåndsdefinerte betydningsskiller i kombinasjon med korpusbaserte metoder.

1.4.2 Korpusbaserte (datadrevne) tilnærminger til WSD

Etter som større tekstkorpora har blitt tilgjengelig i økende grad fra 90-tallet og utover, har mye av forskningen dreid seg om å undersøke mulighetene for korpusbaserte (datadrevne) metoder for WSD. Tekstkorpora består av store tekstsamlinger hentet fra naturlig språk, for eksempel et aviskorpus (med samlinger av tekster fra aviser). Et korpus kan betraktes som en direkte, empirisk kilde for informasjon om hvordan ordbetydninger kontekstuelt benyttes i faktisk språkbruk. Mens kunnskapsdrevne tilnærminger har den fordelen at de har tilgang på et stort og eksplisitt definert inventar av betydningsskiller, er korpusbaserte tilnærmingers store fordel at de supplerer en rik kilde til faktiske eksempler på hvordan ord brukes i kontekst.

Jurafsky & Martin (2000) definerer tre hovedparadigmer for korpusbaserte maskinlæringstilnærminger til WSD, hvor de to siste kan karakteriseres som varianter av prinsippet bak den første tilnærmingen.

1.4.2.1 Tilnærminger med et betydningstagget korpus (overvåket WSD)

Overvåkede metoder for WSD forutsetter et korpus hvor hver forekomst av ordet/ordene som systemet skal lære å disambiguere, på forhånd er korrekt tagget for betydning ut fra et valgt sett av betydningstagger. Input til den anvendte læringsalgoritmen er slike betydningstaggede forekomster. Læringsalgoritmen anvender disse eksempelkontekstene for å utarbeide kunnskap (som regel statistisk basert) om hva som karakteriserer hver ordbetydning. Terminologien som benyttes er å si at algoritmen lærer seg en *klassifikator* for et gitt flertydig ord, og *klassifiserer* nye forekomster (eller *instanser*) med hensyn til betydning.

Denne metoden forutsetter altså et tekstkorpus som inneholder korrekt betydningstaggede, konkrete eksempler på mulige kontekster for hver ordbetydning. Erfaringen innenfor forskning rundt WSD så langt er at dette er paradigmet som synes å ha størst anvendelsespotensial resultatmessig (Diab & Resnik, 2002; Agirre & Martinez, 2001; Escudero et al., 2000). Dette gjelder ikke minst fordi bruken av korpora prinsipielt er anvendbar for å hente ut kunnskap om kontekster fra et spekter av tematiske områder (Leacock & Chodorow, 1998).

Problemet er at semantisk tagging av et korpus per i dag må utføres manuelt, hvilket er både kostbart og tidkrevende. For bedre å se proporsjonene av arbeidet dette ville gi, kan vi tenke oss at et overvåket WSD-system skulle anvendes innenfor et maskinoversettelsessystem. Da ville det være nødvendig med betydningstagget treningsmateriale for alle flertydige ord i oversettelsessystemets leksikon; hvilket ville gi en enorm arbeidsmengde manuelt.[2] I forsøk må som regel treningsmaterialet derfor begrenses til å dekke kun noen få tvetydige ord, og ofte blir det slik at jo flere flertydige ord som skal tagges, jo færre taggede eksempler får man laget for hvert ord. Den begrensede størrelsen på treningskorpuset gjør i sin tur at maskinlæringsalgoritmen får et dårligere statistisk grunnlag for å "lære" ordbetydninger. For eksempel rapporterer Leacock & Chodorow (1998) at jo større treningskorpus som ble gitt som læringsmateriale til deres korpusbaserte klassifikator, jo bedre presterte klassifikatoren.

Selv om overvåket WSD altså resultatmessig anses som den mest lovende tilnærmingen, møter eksperimenter i større skala på en betraktelig "flaskehals" på grunn av manglende betydningstaggede korpora. Problemet som Gale et al. (1992) betegner som "the knowledge acquistition bottleneck" relateres derfor i dag primært til problemet med manglende betydningstaggede korpora for overvåket WSD. Resnik & Yarowsky (1997/2000) gjør opp status som følger:

> *Given the data requirements for supervised learning algorithms and the current paucity of such data, we believe that unsupervised and minimally supervised methods* ["bootstrapping"; undertegnedes kommentar] *offer the primary near-term hope for broad-coverage sense tagging.*

Som Resnik & Yarowsky (1997/2000) her antyder, kan tilnærmingene med uovervåket læring (1.4.2.2 under) og "bootstrapping" (1.4.2.3) betraktes som forsøk på å henholdsvis unngå eller avhjelpe mangelen på betydningstaggede treningskorpora. I det følgende presenteres de to tilnærmingene i sin "opprinnelige" form, før vi deretter går over til den "nyere" tilnærmingen til WSD, "hybrider" i (1.4.3).

1.4.2.2 Tilnærminger med et utagget korpus (uovervåket WSD)

[2] Moon (2000) estimerer at ca. 25 % av det engelske vokabularet er flertydige innenfor samme ordklasse. Dette gir ca 12.000 ord man ville behøve betydningstaggede korpora for!

Siden problemet med overvåkede maskinlæringstilnærminger begrenses av mangelen på semantisk taggede treningskorpora, har det blitt eksperimentert med korpusbaserte tilnærminger som ikke forutsetter at treningskorpuset er korrekt tagget på forhånd.

I sin "opprinnelige" form består uovervåket læring i å presentere maskinlæringsalgoritmen for et utagget treningskorpus, som simpelthen illustrerer *mulige* kontekster for det/de ord som systemet skal lære å disambiguere. Siden eksemplene ikke på forhånd er assosiert med en betydning, må et "rent" system for uovervåket læring på egen hånd utarbeide hvilke betydningsskiller som foreligger for ordet/ordene og hva som kjennetegner hver betydning. Dette foregår enkelt sagt ved at et flertydig ords kontekstord samles i klynger ("clusters") ut fra kontekstuelle likheter mellom eksemplene (Jurafsky & Martin, 2000). Som eksempel kan vi tenke oss et korpus som eksemplifiserer mulige kontekster for det flertydige norske substantivet *tak*. Antagelsen er at eksempler som inneholder sammenfallende kontekstord, f.eks. ord som *bygge, reparere* og *hus*, representerer én og samme betydning av *tak* (f.eks. "hustak"-betydningen). Likeledes vil en annen klynge av sammenfallende kontekstord antas å tilsvare en annen betydning av *tak*. På denne måten blir det automatisk generert et betydningstagget treningskorpus, som så anvendes for å trene en klassifikator etter prinsippet for overvåket WSD.

De observerte problemene med tilnærmingen er at siden betydningsskiller utledes av algoritmen selv, er det vanskelig å evaluere klassifikatorens resultater mot hva skulle være å regne som "riktig" orddisambiguering av gitte ordforekomster: For det første er ikke alltid den uovervåkede metodens egendefinerte betydninger overhodet kjent, for det andre er antallet klynger (ordbetydninger) nesten alltid ulikt antallet ordbetydninger vi ville definert intuitivt eller på basis av en leksikalsk ressurs (Jurafsky & Martin, 2000). Som vi skal se i (1.4.3) under, kombineres derfor uovervåket læring i dag med forhåndsdefinerte betydningsskiller fra en leksikalsk kunnskapsressurs.

1.4.2.3 "Bootstrapping"-tilnærminger til WSD

"Bootstrapping"-tilnærmingen til WSD befinner seg i grenselandet mellom overvåkede og uovervåkede tilnærminger. Det kan bemerkes at Stevenson (2003) inkluderer denne tilnærmingen under kategorien av han kaller "hybrid"-tilnærminger (1.4.3 under). Hans valg bunner i at "bootstrapping", som de uovervåkede metodene, ofte i praksis går på tvers av skillet mellom kunnskapsbaserte og korpusbaserte tilnærminger. Jurafsky & Martin (2000) kategoriserer imidlertid "bootstrapping" som en tilnærming som i sin "rene" form er en kombinasjon mellom overvåkede og uovervåkede korpusbaserte tilnærminger, altså en type korpusbasert tilnærming. Meningen med denne metoden er å *redusere* antallet forhåndtaggede treningseksempler som må utarbeides på forhånd. Resnik & Yarowsky (1997/2000) bruker derfor begrepet *minimally supervised methods* om denne tilnærmingen.

Utgangspunktet for "bootstrapping" er et lite treningssett hvor hver forekomst av et flertydig ord *x* er korrekt betydningstagget. (I utgangspunktet skjer dette manuelt, men som vi vil se i (1.4.3) eksperimenteres det også med å la betydningstaggingen foregå automatisk, basert på betydningsskiller fra en leksikalsk ressurs). Den vanlige terminologien er å omtale dette korpuset av forhåndstaggede eksempelforekomster av hver betydning som engelsk *seeds*. Dette kan vi på norsk kalle *frø*, eller kanskje bedre "bootstrapping"-frø.

Basert på et forholdsvis lite antall av slike "bootstrapping"-frø som eksemplifiserer bruken av hver ordbetydning, ekstraherer systemet selv nye eksempler på kontekst for hver betydning. Dette foregår ved at systemet trener en første klassifikator (ved overvåket læring) basert på "bootstrapping"-frøene. Denne klassifikatoren disambiguerer dernest alle nye

ordforekomster i korpuset som inngår i en kontekst som klassifikatoren med stor sikkerhet gjenkjenner ut fra "bootstrapping"-frøene. Forekomster med kontekster som klassifikatoren ennå ikke er sikker på, utsettes. Med de nye forekomstene lagt til i treningskorpuset, trenes en ny klassifikator på dette utvidede korpuset, og denne prosedyren gjentas iterativt slik at treningskorpus blir stadig større og stadig sikrere klassifikatorer blir trent. Når hele det opprinnelig utaggede korpuset er betydningstagget (eller så mange forekomster som "bootstrapping"-algoritmen greier å tagge), anvendes dette korpuset for å trene en siste klassifikator etter prinsippet for overvåket WSD.

1.4.2.4 Oppsummering

Korpusbaserte tilnærmingers store fordel, til forskjell fra kunnskapsbasert WSD, er at et tekstkorpus gir rik informasjon om ordbetydningers mulige kontekster. Problemet med tilnærmingen er at tilgjengelige tekstkorpora (som for eksempel et aviskorpus) i utgangspunktet bare er samlinger av tekst som ikke er tagget for betydning. For at læringsalgoritmen skal kunne assosiere et gitt ord i kontekst med en viss ordbetydning, må korpuset dermed "eksplisitt" tagges for betydning på forhånd, hvilket i utgangspunktet må foregå manuelt (overvåket læring).

På grunn av det omfattende og tidkrevende arbeidet med å utføre betydningstaggingen av et korpus manuelt, har det blitt forsøkt å la læringsalgoritmen selv utarbeide hvilke ordbetydninger som foreligger på basis av mulige kontekster for et ord (uovervåket læring), eller å redusere mengden av nødvendig forhåndstagget materiale ("bootstrapping"). Uovervåket uovervåket læring er problematisk fordi det blir vanskelig å kontrollere hvilke betydningsskiller klassifikatoren utleder, mens "bootstrapping" bare begrenser mangelen på betydningstagget materiale for hver ordbetydning uten å virkelig unngå problemet.

Dette leder oss i (1.4.3) under til nyere forskning, hvor uovervåket WSD og "bootstrapping" kombineres med leksikalske ressurser assosiert med de kunnskapsbaserte WSD-tilnærmingene.

1.4.3 "Hybrid"-tilnærminger for WSD

I Jurafsky & Martin (2000) sin oversikt over tilnærminger til WSD defineres ikke en egen "hybrid"-kategori for WSD-eksperimenter som går på tvers av det per i dag "etablerte" skillet mellom kunnskapsdrevet WSD og korpusbasert WSD. Det synes likevel fornuftig av Stevenson (2003) å foreslå introduksjonen av et eget begrep for slike tilnærminger: Vi har sett at de leksikalske ressursenes fordel er at de gir eksplisitt informasjon om betydningsskiller (1.4.1), mens korpora på sin side gir rik kontekstuell informasjon (1.4.2). Det kan derfor anses som en naturlig utvikling at mye forskning innenfor WSD i dag eksperimenterer med å kombinere de to ressursene.

Typisk for "hybrider" er at metoden, som uoveråket WSD og "bootstrapping", grunnleggende sett retter seg mot problemet med manglende tilgang på betydningstaggede korpora: Steg én innebærer å undersøke kombinasjonsmuligheter mellom leksikalske kunnskapsressurser og uovervåket læring/"bootstrapping" for å automatisere betydningstaggingen av et treningskorpus. Det resulterende treningskorpuset anvendes i steg to for å trene en WSD-klassifikator etter prinsippet for overvåket WSD.

Yarowsky (1995) presenterer for eksempel en "bootstrapping"-metode for WSD hvor også de såkalte frøene for "bootstrapping" kan genereres automatisk. Metoden tar utgangspunkt i forhåndsdefinerte betydningsskiller for de aktuelle ordene, og i første steg

deriveres et lite betydningstagget treningssett (frø) automatisk for hver av betydningene til et flertydig ord. Dette gjøres ved å definere "sikre" kollokasjonsord, det vil si kontekstord i posisjoner umiddelbart rundt den flertydige ordforekomsten som kan antas å være sikre indikatorer for hver betydning. Yarowsky gir som eksempel at kollokasjonen *manufacturing plant* kan regnes som en sikker indikator på at engelsk *plant* da har betydningen "anlegg/fabrikk", mens *plant life* utgjør en sikker kollokasjonsindikator på at det dreier seg om *plant*|"plante". (Yarowsky foreslår å hente slike kollokasjonsindikatorer fra WordNet, ordboksdefinisjoner eller å utarbeide dem manuelt).

Alle forekomster av *plant* i korpus som har disse "sikre" kollokasjonsordene tagges automatisk for betydning, og utgjør således frø for en "bootstrapping"-prosedyre. Først trenes en klassifikator på frøene, og denne disambiguerer (i vårt eksempel) nye forekomster av *plant* når klassifikatoren er tilstrekkelig "sikker" ifølge en definert terskel. Alle "sikre" nye forekomster blir lagt til i det betydningstaggede korpuset, og på grunnlag av dette utvidede korpuset trenes en ny klassifikator. Slik trenes iterativt nye klassifikatorer for å utvide treningsmaterialet, frem til hele korpuset er disambiguert og en siste klassifikator for WSD trenes på hele dette betydningstaggede korpuset. Testmaterialet for den endelige klassifikatoren besto av allerede taggede semantiske korpora fra tidligere eksperimenter. Med denne uovervåkede/"bootstrapping"-metoden rapporterer Yarowsky at systemet gjennomsnittlig greier å disambiguere ny tekst med en presisjon på 96,1 % (basert på de 12 substantivene som inngikk i eksperimentet).

Det har blitt utført flere interessante forsøk som kombinerer informasjon fra WordNet (1.4.1.3) med korpusbasert læring (bl.a. Resnik, 1995 og Leacock & Chodorow, 1998). Mest relevant for denne hovedoppgaven er et eksperiment utført av Chodorow et al. (1998), hvor et treningskorpus for flertydige ord blir generert automatisk ut fra WordNets betydningsskiller og semantiske relasjoner. Som vi husker fra (1.4.1.3), er substantiver i WordNet organisert i hierarkiske sett av synonymer (kalt *synsets*) som lenkes sammen ved semantiske relasjoner som hyperomymi/hyponymi (over- og underbegreper), antonymi og meronymi. Tanken bak dette eksperimentet er at entydige ord som er semantisk nærbeslektet med det aktuelle flertydige ordets betydninger i WordNet, kan antas å forekomme i kontekster som er "nærbeslektet" med selve det flertydige ordets mulige kontekster. Eksperimentet består derfor i å hente ut ikke-flertydige ord i WordNet som er nærbeslektet med hver betydning av det aktuelle flertydige ordet, og å bygge et treningskorpus som inneholder de *nærbeslektede* ordene framfor selve det flertydige ordet.

Siden WordNet bare har informasjon om hyperomymi/hyponymi mellom substantiver, var eksperimentet i praksis begrenset til å generere treningskorpora kun for substantiver. Forsøket tok utgangspunkt i et lite knippe flertydige substantiver (14 stk.) og deres betydningsdistinksjoner i følge WordNet. Som eksempel er engelsk *line* oppført med fem betydninger i WordNet. For hver av disse fem betydningene blir deres 100 nærmest relaterte entydige ord hentet ut fra WordNet[3]. Når hver betydning av *line* er tildelt sine respektive 100 entydige og nærrelaterte ord, søkes det i et utagget tekstkorpus etter eksempelsetninger som inneholder hver *line*-betydnings nærrelaterte ord. Siden de utvalgte relaterte ordene er entydige, er sannsynligheten liten for at metoden feilaktig henter ut eksempelsetninger fra korpuset som ikke på noen måte har med *line*s aktuelle ordbetydninger å gjøre.

På denne måten ekstraheres automatisk, eller "uovervåket", et treningskorpus for hver av de 14 ordene som inngikk i eksperimentet. Chodorow et al. foreslår å kalle et slikt korpus for *monosemic relative corpus*, som vi på norsk kan oversette som et *entydig slektning-*

[3] Entydige, nærrelaterte ord velges i følgende rekkefølge: Synonymer, direkte underordnede datterkonsepters ord som har *line* som hode (for eksempel *trap line*), deretter øvrige underbegreper, hyponymer som inneholder *line* (uten at *line* må være hode i sammensetningen), alle andre hyponymer; hypernymer (overbegrep) og til slutt søsterkonsepter.

korpus. En WSD-klassifikator trenes så etter overvåket WSD-prinsippet på slike *entydig slektning*-korpora, og testes deretter på forekomster av de faktiske flertydige ordene. De 14 *entydig slektning*-klassifikatorenes prestasjonsnivå ble så sammenlignet med resultatene fra å trene en klassifikator basert på et *manuelt* betydningstagget treningskorpus for de 14 ordene som inngikk i eksperimentet. Chodorow et al. rapporterer at de fleste av testordenes respektive *entydig slektning*-korpus presterer 1-2 % svakere enn ved trening på et manuelt betydningstagget korpus. De observerte at for enkelte av ordene var det et problem at *entydig slektning*-korpusets treningseksempler hadde en ujevn fordeling mellom betydningene, slik at enkelte betydninger av et ord hadde svært få treningseksempler. En mer prinsipiell svakhet ved tilnærmingen deres er at klassifikatoren altså ikke trenes på selve det flertydige ordets mulige kontekster. Dette kan få konsekvenser for læringen, siden det flertydige ordets entydige slektninger kan vise seg å inngå i andre kollokasjoner (den umiddelbare, lokale konteksten rundt ordet) enn selve det flertydige ordet. Som eksempel viser de til at "linje/rekke"-betydningen av engelsk *line* ofte er fulgt av en genitivsfrase som i *a line of children*, mens betydningens entydige slektning *picket line* ("streikevakt-linje/sperre") som regel ikke har denne kollokasjonen. Følgelig trenes *entydig slektning*-klassifikatoren på kollokasjoner som ikke nødvendigvis er karakteristisk for selve det flertydige ordet som den etterpå skal klassifisere.

En annen interessant ressurs for automatisk betydningstagging av et korpus er parallellkorpora. Den grunnleggende observasjonen bak ideen er at ulike betydninger av et flertydig ord ofte oversettes ulikt i et annet språk (Brown et al., 1991). For eksempel kan det norske substantivet *tak* oversettes til engelsk *roof* i "hustak"-betydningen, og som *grip* i "gripetak"-betydningen. Et relevant forsøk av nyere dato er gjennomført av Diab & Resnik (2002). Deres presenterte metode anvender oversettelseskorrespondanser i et parallellkorpus og WordNets betydningsskiller som kunnskapsressurser for å utføre uovervåket automatisk betydningstagging av begge sider av parallellkorpuset. Metoden forutsetter et parallellkorpus som er sidestilt på ordnivå for å kunne hente ut oversettelseskorrespondanser. I Diab & Resniks eksperiment anvendte de et kunstig parallellkorpus, skapt ved å la et engelskspråklig korpus bli oversatt til fransk ved hjelp av maskinoversettelsessystemer.

Metodens første steg er å starte med ord på den ene siden av parallellkorpuset, og registrere ordenes oversettelseskorrespondansene på den andre siden av parallellkorpuset. Som eksempel kan vi si at prosessen begynner med å registrere engelske ords oversettelseskorrespondanser i den franske siden av parallellkorpuset. Hver ordforekomst det registreres en oversettelse for blir registrert med hensyn til dens posisjon i korpus, slik at hver forekomst til slutt kan betydningstagges automatisk.

Neste steg er å gå til den motstående siden av parallellkorpuset, i vårt tilfelle fransk, og samle sammen hvert franske ord *x*s sett av engelske oversettelseskorrespondanser: $\{y_1, y_2, \ldots, y_n\}$. Diab & Resnik bruker som konkret eksempel at fransk *tragedy* oversettelsesmessig er funnet å være assosiert med engelsk {*tragedy*, *disaster*, *situation*}. Slike sett refereres til som et *target set* av oversettelseskorrespondanser. I det påfølgende steget skal hvert slikt *target set* assosieres med en WordNet-betydning. Den bakenforliggende intuisjonen er at selv om ordene i et *target set* kan være flertydige, favoriserer forekomsten av for eksempel *disaster* og *tragedy* i samme *target set* et betydningsaspekt som ordene har *felles*. Dette felles betydningsaspektet utarbeides med utgangspunkt i WordNets mulige ordbetydninger for hvert av ordene i et *target set*. Alle ordpar $<y_i, y_j>$ (j≠i) i et *target set* sammenlignes etter tur mot WordNets betydningsinventar for de aktuelle ordene, og for hvert ordpar blir det kalkulert hvilken WordNet-betydning som semantisk er nærmest begge ord i et par. (Diab & Resnik gir som eksempel at *tragedy* sannsynligvis vil stå semantisk nærmere både *disaster* og *situation* i sin WordNet-betydning *tragedy*|"calamity", framfor i *tragedy*|"kind-of-drama"-betydningen.) Hver "parvise" betydning tildeles dessuten en konfidensverdi som reflekterer hvor nært de to

ordene i et par var beslektet. Det endelige valget av betydning for et *target set* gjøres ved å velge den ordbetydningen som hadde størst konfidensverdi av alle. (I eksempelet ovenfor: CALAMITY.)

Når et *target set* så er tildelt den mest sannsynlige WordNet-betydningen, er neste steg å utføre selve betydningstaggingen av korpuset. Siden første steg registrerte posisjonen til alle ord som inngikk i uthentingen av oversettelseskorrespondanser, kan i eksempelet ovenfor nå alle forekomster av engelsk *tragedy, disaster* og *situation* som er assosiert med fransk *tragedy* tagges som CALAMITY. Likeledes kan man i neste steg overføre den engelske sidens betydningstagger til de parallellstilte ordene på den franske siden.

Diab & Resnik observerer at metoden for automatisk betydningstagging ikke er anvendbar for de ord som maskinoversettelsessystemet konsekvent velger samme oversettelse for, siden metoden avhenger av å kunne måle semantisk likhet mellom to oversettelseskorrespondanser i et *target set*. De påpeker derfor at et parallellkorpus med menneskelige oversettelser antagelig vil være bedre egnet for å få større variasjon i settene av oversettelseskorrespondanser. En videre begrensning på hvilke forekomster av et ord som kan betydningstagges, er at forekomsten må inngå i en setningen som er parallellstilt med en setning i det motstående språket[4]. Med hensyn til presisjon (hvor mange av de taggede forekomstene som er korrekte) finner de at det største problemet bunner i "støy" (feil) i ordparallellstillingen, slik at uønskede ord feilaktig inkluderes i settene av oversettelseskorrespondanser og påvirker målet på semantisk likhet innbyrdes i et *target set*. Metodens kanskje største svakhet er at den implisitt forutsetter at alle de kildeordene som gir opphav til et *target set* av oversettelseskorrespondanser skal representere samme betydning. Hvis kildeordene representerer ulike betydninger, vil det resulterende *target set* inneholde oversettelser som reflekterer flere mulige og urelaterte betydninger av de aktuelle kildeordene. Åpenbart vil dette gi konsekvenser for beregningen av hvilken WordNet-betydning som semantisk står nærmest ordene i et *target set*.

Diab & Resnik slår likevel fast at denne uovervåkede metoden for automatisk betydningstagging av et korpus resultatmessig er på høyde med andre uovervåkede tilnærminger. De påpeker at selv om man ikke kan forvente at et uovervåket system presterer like godt som en overvåket WSD-tilnærming, burde metoden likevel være anvendbar som såkalte frø i en "bootstrapping"-prosedyre for å tagge større mengder av korpusdata.

1.4.4 Diskusjon

Det er verdt å merke seg at WSD-tilnærminger vanligvis kategoriseres i henhold til hvilke bakenforliggende ressurser som legges til grunn (som i presentasjonen ovenfor). Et poeng som omtales lite i litteraturen, men som kan være nyttig å være klar over, er imidlertid at om man i stedet kategoriserte WSD-metoder i henhold til anvendelsesformål, ville det være naturlig å skille mellom to formål: (I) Metoder som prinsipielt sikter direkte mot leksikalsk flertydighet som problem innenfor NLP-systemer, og (II) metoder som retter seg mot betydningstagging av et finitt treningskorpus som *bakenforliggende* ressurs for WSD.

Som vi husker fra (1.3), blir det sagt at det overordnede målet for WSD er å komme frem til metoder som kan benyttes for disambiguering i stor skala innenfor "større" NLP-systemer, f.eks. maskinoversettelsessystemer og automatisk sammendrag av tekst (Wilks & Stevenson, 1996). Dette innebærer at metoden prinsipielt er forventet å kunne disambiguere en potensielt infinitt mengde av nye forekomster av et gitt flertydig ord. Alle tilnærminger

[4] Det kunstige parallellkorpuset som ble brukt i Diab & Resniks eksperiment ble ifølge artikkelen automatisk ordparallellstilt ved hjelp av en GIZA++-implementering av IBMs statistiske Maskinoversettelsesmodeller. Som referanse oppgir de Och & Ney (2000), referansen er inkludert i denne oppgavens referanseliste.

presentert ovenfor; kunnskapsbaserte, korpusbaserte og "hybrid"-tilnærminger, kan sies å være metoder hvis endelige mål er å være anvendbare for bruk innenfor "større" NLP-systemer.

Imidlertid har vi også sett at den manglende tilgangen på semantisk taggede treningskorpora for overvåket WSD leder til WSD-metoder hvor en egen form for orddisambiguering inngår som et *delmål* (II): Forsøkene presentert i (1.4.3) eksemplifiserer automatisk orddisambiguering som foregår med det spesifikke delmålet å ekstrahere finitte, betydningstaggede treningskorpora som *bakenforliggende* ressurs for WSD av måltype (I). For denne typen orddisambiguering kunne man formulere som eksplisitt mål at metoden skal generere et finitt, betydningstagget korpus, og fortrinnsvis skal metoden prinsipielt være egnet for å generere slike finitte treningskorpora for hele åpen ordklasse-vokabularet i et språk.

For eksempel så vi i (1.4.3) at Diab & Resnik (2002) bruker oversettelseskorrespondanser i et parallellkorpus som ressurs for automatisk betydningstagging. Det ligger da i metodens natur at antallet ordforekomster som kan tagges vil være begrenset av hvorvidt det var mulig å identifisere en parallellstilt oversettelse i det motstående språket for hver gitte ordforekomst. Følgelig kan vi tolke metoden som at den evner å ekstrahere et finitt, betydningstagget korpus, men noen forekomster i et opprinnelig utagget korpus vil ganske enkelt forbli utagget. Forutsatt at der foreligger oversettelseskorrespondanser i et parallellkorpus over et stort vokabular, kan metoden prinsipielt forventes å greie å generere finitte treningskorpora for alle flertydige ord i vokabularet. Likeledes så vi i Chodorow et al. (1998) sitt eksperiment at uovervåket ekstrahering av et finitt treningskorpus per definisjon ikke disambiguerer selve det flertydige ordet, men henter ut ordets entydige "slektninger" i WordNet. Følgelig er heller ikke denne metoden å regne som en form for orddisambiguering som er direkte anvendbar for WSD som hovedmål. Imidlertid synes metoden anvendbar for delmålet med å ekstrahere en form for indirekte betydningstagget treningskorpus.

Poenget er at det kan være nyttig å være klar over at WSD som hovedmål kan sies å legge andre forventninger til grunn enn delmålet. (potensielt infinitte mengder vs. finitte mengder av betydningstaggede ord).

I neste delkapittel presenteres hypotesen og metoden bak denne hovedoppgaven. Som vi vil se, vil oppgaven defineres som å ha to deler: Hovedvekten legges på en metode for automatisk ekstrahering av et betydningstagget, finitt korpus i del én, mens del to utgjør et forsøk på å trene en klassifikator for WSD rettet mot det opprinnelige målet.

1.5 Speilmetoden og parallellkorpus som ressurs for automatisk ekstrahering av et betydningstagget korpus for WSD

Denne oppgaven kan karakteriseres som en "hybrid"-tilnærming til WSD, og består av to deler. Del én (1.5.1 under) er rettet mot korpusbasert WSDs mangel på betydningstaggede korpora, og foreslår en uovervåket metode som ekstraherer et finitt, betydningstagget korpus på basis av oversettelseskorrespondanser i et parallellkorpus. Vi vil referere til metoden som en metode for *Automatisk Betydningstagging av Treningskorpus*, som også kan forkortes til ABT. Del to av oppgaven (1.5.2 under) vil bestå i å trene en WSD-klassifikator basert på det resulterende automatisk betydningstaggede treningskorpuset. Hovedfokus vil legges på ABT-metoden presentert i del én, og det primære målet med del to er å benytte treningen av en klassifikator for å evaluere ABT-materialet. I det følgende konkretiseres hypotesen og mål for oppgaven.

1.5.1 Automatisk Betydningstagging av et Treningskorpus (ABT)

ABT-metodens bakenforliggende nøkkelressurs er en metode utviklet av Helge Dyvik (Dyvik, 1998/2002) innenfor forskningsprosjektet "Fra Parallellkorpus til Ordnett" (finansiert av Meltzerfondet og Norges forskningsråd). Dyviks speilmetode anvender oversettelseskorrespondanser i et parallellkorpus for å utlede semantiske egenskaper ved ord automatisk, deriblant et ords flertydighet i forhold til et annet språk[5]. I dette eksperimentet anvendes English-Norwegian Parallel Corpus (ENPC), utviklet i samarbeid mellom Universitetet i Oslo og HIT-senteret i Bergen[6].

Hypotesen bak speilmetoden er at kontrastivt flertydige ord ikke forventes å ha oversettelser med samme flertydighet, og at man heller ikke venter at flere ord i samme språk skal ha den samme kontrastive flertydigheten. Med disse antagelsene som utgangspunkt definerer speilmetoden et gitt ord *x*s betydningsskiller i forhold til et motstående språk, ut fra hvordan *x*s oversettelseskorrespondanser overlapper for andre ord enn *x*.

Som et forenklet eksempel kan vi tenke oss at det norske substantivet *rett* ifølge et parallellkorpus viste seg å ha følgende sett av oversettelsespartnere i engelsk: {*dish*, *food*, *right* og *entitlement*}. Intuitivt aner vi av oversettelseskorrespondansene at substantivet *rett* har minst to betydninger som vi kan kalle *a* og *b*. Siden vi ikke venter at andre ord i norsk eller det motstående språket har den samme flertydigheten som *rett*, venter vi at ord med betydning *a* og ord med betydning *b* ikke skal ha noen felles oversettelser utover ordet *rett* selv. Derimot venter vi at *dish* og *food* kan ha andre felles oversettelser enn *rett*, og dette utgjør et kriterium på at de to engelske ordene er semantisk beslektet. Det samme gjelder paret *right* og *entitlement*. Hvis vi så antar at semantisk beslektede ord normalt har mer enn en oversettelseskorrespondanse felles, blir konklusjonen at henholdsvis *dish/food* og *right/entitlement* representerer to ulike betydninger av *rett*. Siden vi antar at hvert slikt subsett reflekterer en av *rett*s betydninger, refererer vi til subsettene som *betydningspartisjonene* til *rett*s sett av oversettelseskorrespondanser.

Disse betydningspartisjonene utgjør utgangspunktet for den foreslåtte ABT-metoden. Som påpekt av bl.a. Dyvik (1998/2002) og Diab & Resnik (2002), kan oversettelsesdata i et parallellkorpus betraktes som resultatet av en prosess hvor det oversatte ordet har blitt tolket i sin kontekst. Hypotesen som legges til grunn for ABT-metoden som presenteres her er dermed at hvis speilmetodens oversettelsesmessige betydningspartisjoner for et ord *x* reflekterer ulike betydninger av *x* i forhold til et annet språk, så kan hver betydningspartisjon tolkes som (et korrelat til) en semantisk kategori av *x*. I eksempelet ovenfor ville *rett*N tolkes

[5] En grensesnitt for anvendelse av speilmetoden er implementert i LISP Medley og er tilgjengelig fra: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorsguide.html

[6] Informasjon om ENPC og tilgang til korpuset fra: http://www.hf.uio.no/iba/prosjekt/

som å ha to semantiske kategorier. ABT-metoden tar således et ord *x*s betydningspartisjoner (ved speilmetoden) som input, og tagger forekomster av *x* i parallellkorpuset semantisk ut fra hvilken semantisk kategori forekomstens oversettelse er tilordnet: Hvis en forekomst av *rett*N er funnet å korrespondere med *dish* i en gitt setning i parallellkorpuset, tagges denne forekomsten med den semantiske kategorien (dvs. betydningspartisjonen) som inneholder *dish*; hvis en forekomst korresponderer med *entitlement* tilordnes forekomsten *entitlement*s semantiske kategori, osv. På denne måten kan et subsett av parallellkorpuset konverteres til et monolingualt korpus som er tagget for betydning – I eksempelet ovenfor vil dette subsettet bestå av semantisk taggede forekomster av substantivet *rett*N i kontekst.

Del éns forsøk på å implementere en automatisk betydningstagger er motivert av problemet med manglende tilgang på semantisk taggede treningskorpora til bruk innen maskinlæringsmetoder for WSD. Formålet med forsøket er å undersøke i hvilken grad oversettelseskorrespondanser i parallellkorpus, sortert etter betydning ved speilmetoden, er egnet som ressurser for å automatisere betydningstaggingen av et treningskorpus. I (1.4.3) så vi at Diab & Resnik (2002) har utført et lignende eksperiment. I deres eksperiment blir et ord *x*s sett av oversettelseskorrespondanser $\{y_1, y_2, \ldots, y_n\}$ tilordnet den betydningen fra WordNets begrepshierarki som semantisk ligger nærmest ordene $\{y_1, y_2, \ldots, y_n\}$. Med andre ord anvender de WordNets manuelt definerte betydningsskiller, og kombinerer disse med informasjon om oversettelseskorrespondanser i et parallellkorpus. Vi så videre som en svakhet ved deres metode at den implisitt forutsetter at *x* må være entydig, siden metoden ikke tar høyde for at settet $\{y_1, y_2, \ldots, y_n\}$ kan inneholde ord som innbyrdes er semantisk urelaterte.

Til forskjell fra deres tilnærming tar denne oppgavens ABT-metode utgangspunkt i betydningsskiller som er derivert direkte ut fra oversettelseskorrespondanser ved speilmetoden (framfor å være manuelt definert). Speilmetoden tar dessuten høyde for flertydighet hos ethvert ord som det registreres oversettelseskorrespondanser for: Poenget med speilmetoden er at den baserer seg på overlapping mellom *flere* ords sett av oversettelseskorrespondanser (framfor å måle alle ord innenfor ett sett mot en ekstern kunnskapsbase). Innenfor forskningsprosjektet "Fra Parallellkorpus til Ordnett" (se 2.2) er speilmetoden hovedsaklig et verktøy for å generere et semantisk ordnett. Betydningsskillene som speilmetoden utleder for et gitt lemma har imidlertid ikke blitt testet i en praktisk orddisambigueringsoppgave (WSD). Siden det automatisk betydningstaggede materialet i denne oppgaven vil testes i en maskinlæringsalgoritme for WSD, er det derfor interessant å evaluere hvorvidt speilmetodens deriverte betydningskiller synes anvendbare i en praktisk oppgave som WSD.

Det praktiske arbeidet i del én er todelt. Første steg er å anvende speilmetoden for å hente ut et gitt ords oversettelsesbaserte betydningsskiller. Ordparallellstilling av parallellkorpuset ENPC er under utvikling, men da ENPC per i dag bare er parallellstilt på setningsnivå, må oversettelseskorrespondanser hentes ut manuelt. Speilmetodens utledede betydningsskiller for de to valgte lemmaene vil evalueres mot betydningsskillene for de samme to lemmaer ifølge Bokmålsordboka. Som vi skal se, ble bare ett av lemmaene funnet å være egnet som testlemma for automatisk betydningstagging.

Den andre delen består i å implementere selve den automatiske betydningstaggeren. Metoden er implementert i LISP Allegro. Evalueringen av ABT-metoden vil foregå på to måter: Innenfor del én testes metoden på betydningsskillene som speilmetoden utledet for testlemmaet *rett*N. Det resulterende betydningstaggede treningskorpuset for *rett*N vil så gis en evaluering basert på manuell gjennomgang med henblikk på *presisjon* (antall forekomster som ble korrekt betydningstagget) og *recall* (hvor mange forekomster som ble forsøkt tagget). Spørsmålet her vil primært være hvilke feil som forekommer i materialet: Bunner de i en feilkilde i speilmetoden (betydningspartisjonene) eller i andre feil som kan relateres til selve prinsippet bak ABT-metoden? Alternativet vil være at eventuelle feil kommer av problemer

av mer praktisk art, f.eks. fordi parallellkorpuset per i dag ikke er parallellstilt på ordnivå. Del to av oppgaven kan betraktes som en praktisk evaluering av ABT-metoden.

1.5.2 Trening av en overvåket WSD-klassifikator

Et naturlig neste steg er å trene en klassifikator på ABT-metodens betydningstaggede treningskorpus etter prinsippet for overvåket WSD. Det ble valgt å benytte en læringsalgoritme fra software-pakken TiMBL. TiMBLs input må være kodet som en trekkvektor med et fast antall trekk for hvert treningseksempel. Dette betyr at alle treningsforekomster av ordet som skal disambigueres må ha et likt antall *n* kontekstord.

Første steg for å trene en klassifikator var derfor å avgjøre hvilken type kontekst som skal ekstraheres fra korpuset for hver ordforekomst. Innenfor rammene av denne oppgaven ble det hovedsakelig valgt å basere seg på Chodorow et al. (1998) og Leacock & Chodorow (1998). De har gjennomført systematiske eksperimenter på bruken av lokal, posisjonsspesifikk kontekst rundt det aktuelle ordet versus bruk av nøkkelord (åpen klasse-ord i et større vindu), samt en kombinasjon av disse. For substantiver, som er relevant for denne oppgaven, antyder de tidligere eksperimentene at nøkkelord synes å være noe mer informativt enn lokal kontekst for substantiver; men aller best fungerer en kombinasjon av begge typer informasjon. Imidlertid har eksperimenter vist at en klassifikators prestasjoner har en tendens til å variere fra ord til ord, bl.a. avhengig av hvilke betydningsskiller som er tilknyttet det gitte ordet (Chodorow et al, 1998; Leacock & Chodorow, 1998; Hoste et al., 2001). Det synes derfor formålstjenlig å kartlegge egenskapene til testlemmaet innenfor denne oppgaven med hensyn til hvilke kontekstuell informasjon som fungerer bedre, og å anvende denne typen informasjon i de påfølgende eksperimenter for å evaluere ABT-materialet.

Hovedvekten i del to legges på å trene en endelig klassifikator (med en avgjort type kontekstuell informasjon) på det automatisk betydningstaggede materialet fra del én. Formålet er å undersøke hvorvidt ABT-materialet egner seg for trening av en klassifikator. Det naturlige sammenligningsgrunnlaget for en slik evaluering er å måle hvilken kvalitet det resulterende korpuset vil vise seg å ha i forhold til et tilsvarende treningskorpus som er *manuelt* tagget for betydning, både med hensyn til "støy" (feilkilder) i treningsmaterialet som ABT-metoden eventuelt medfører og med tanke på størrelse.

Evalueringen av dette foregår ved først å trene en klassifikator K1 på basis av ABT-metodens materiale for testlemmaet. K1 er således ment å antyde hvor godt maskinlæringsalgoritmen presterer ved bruk av et automatisk ekstrahert treningskorpus, slik metoden for ABT er implementert i denne oppgaven og med parallellkorpuset ENPC som testressurs. Dernest korrigeres ABT-materialet for feil som er av praktisk snarere enn metodisk natur, og en klassifikator K2 trenes ut fra dette delvis korrigerte materialet. K2 kan dermed betraktes som en indikator på hvordan selve metoden for ABT presterer, under "ideelle" implementeringsmessige omgivelser for øvrig. Til sist trenes en klassifikator K3 på grunnlag av et treningskorpus som er manuelt tagget med speilmetodens betydninger for testlemmaet. Denne klassifikatoren kan betraktes som en "baseline" eller "gullstandard" for sammenligning.

Resten av denne oppgaven er organisert som følger:

Kapittel 2 presenterer de to ressursene som skal anvendes, parallellkorpuset ENPC og speilmetoden. Siden speilmetoden er den sentrale bakenforliggende ressursen, vil hovedfokus legges på denne. Kapittelet vil omfatte det manuelle arbeidet med å hente ut oversettelseskorrespondanser manuelt, og forklare hvordan speilmetoden utleder betydningsskiller. Speilmetodens resultater for de to valgte lemmaene vil deretter evalueres.

Kapittel 3 presenterer implementeringen av metoden for å ekstrahere et finitt, betydningstagget treningskorpus, og gir en manuelt basert evaluering av det resulterende materialet.

Kapittel 4 beskriver (I) Eksperimenter for hvilken type kontekstuell informasjon som synes best egnet for det aktuelle testlemmaet, og (II) treningen av tre klassifikatorer som beskrevet ovenfor, for å evaluere det automatisk betydningstaggede materialet.

Kapittel 5 diskuterer oppgavens resultater i forhold til problemstillingen med mangelen på betydningstaggede korpora og i forhold til potensial for videreutvikling.

Kapittel 6 inneholder oppsummering og konklusjon.

2. Oversettelsesbaserte betydningsdistinksjoner. Parallellkorpuset ENPC og speilmetoden

2.1 Innledning

Dette kapittelet presenterer de bakenforliggende ressursene til den presenterte metoden for automatisk ekstrahering av et betydningstagget korpus: Parallellkorpuset ENPC[7] og speilmetoden[8] (Dyvik 1998, 2002). Av disse er speilmetoden den sentrale ressursen, som anvender oversettelseskorrespondanser i et parallellkorpus for å ekstrahere informasjon om et ords flertydighet i forhold til et annet språk.

Som forklart i (1.5) er denne hovedoppgavens primære mål å anvende speilmetodens oversettelsesbaserte betydningsdistinksjoner som ressurs for å betydningstagge forekomster av et flertydig ord i parallellkorpus. Oppgavens første steg er derfor å la speilmetoden utlede et ords betydningsdistinksjoner. Rent praktisk betyr dette å aller først supplere speilmetoden med dens nødvendige oversettelseskorrespondanser i tilknytning til det aktuelle ordet. Da denne oppgavens anvendte korpus, ENPC, foreløpig bare er parallellstilt på setningsnivå, må oversettelseskorrespondanser på ordnivå ekstraheres manuelt fra korpuset. Som en følge av tidsbruken som dette manuelle arbeidet innebærer, ble det valgt å begrense seg til ett flertydig og rimelig frekvent lemma i ENPC. Valget falt på *rett*, som enten kan være substantivlemma (fra nå av referert til som *rett*N) eller adjektivlemma (fra nå av: *rett*A).

Delkapittel (2.2) presenterer parallellkorpuset ENPC. (2.3) tar for seg speilmetoden, og er organisert som følger: Etter en introduksjon av speilmetoden følger en forklaring og diskusjon av prinsippene som ble lagt til grunn for å ekserpere oversettelseskorrespondanser manuelt. Deretter gis et mer detaljert innblikk i de sider ved speilmetoden som er relevante for denne oppgaven, det vil si selve prosedyren med å "speile" ord frem og tilbake oversettelsesmessig (illustrert ved *rett*N). I del (2.4) presenteres speilmetodens resulterende betydningsskiller for testlemmaene *rett*N (2.4.1) og *rett*A (2.4.2). Resultatene evalueres mot den nettbaserte Bokmålsordboka utviklet ved Universitetet i Oslo. I del (2.5) følger en diskusjon og konklusjon med hensyn til speilmetoden.

[7] Tilgjengelig fra: http://www.hf.uio.no/iba/prosjekt/
[8] Tilgjengelig fra: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorsguide.html

2.2 English-Norwegian Parallel Corpus (ENPC)

Parallellkorpuset ENPC[9] er utviklet i samarbeid mellom Universitetet i Oslo og HIT-senteret i Bergen (Johansson et al., 1999/2002). ENPC inneholder originaltekster og deres oversettelser, fra engelsk til norsk og fra norsk til engelsk. Korpuset omfatter både skjønnlitterær tekst (30 originaltekster for hvert språk samt deres tilhørende oversettelser) og generell sakprosa (20 originaltekster for hvert språk samt oversettelser). Til sammen utgjør tekstsamlingen ca. 2,6 mill. ord. Foreløpig er korpuset parallellstilt på setningsnivå, men parallellstilling på ordnivå er under utvikling innenfor forskningsprosjektet "Fra Parallellkorpus til Ordnett" (se 2.3 under).

Korpuset er lemmatisert og tagget for part-of-speech (pos). Taggingen av den norske siden av korpus er basert på regelsettet for Oslo-Bergen-taggeren (Bondi Johannesen, 1998). Da Oslo-Bergen-taggeren ikke har et regelsett for engelsk, er den engelske siden av korpuset tagget med regelsettet fra Penn TreeTagger (Santorini 1991).

Som vi vil se i forbindelse med implementeringen av metoden for automatisk betydningstagging, viste det seg at preprosesseringen (lemmatisering og pos-tagging) ikke kom særlig heldig ut for de norske lemmaene denne oppgaven tar for seg, *rett*N og *rett*A. Dette er kanskje ikke overraskende, siden disse oppslagsordene er sammenfallende. Det later imidlertid ikke til at problemene med å disambiguere disse to lemmaene syntaktisk er representative for Oslo-Bergen-taggerens generelle prestasjonsnivå. Evalueringsresultater fra 2002[10] viser at på bokmålssiden har taggeren en *leksikalsk funnrate* (*recall*) på 99 %, dvs. at den beholder 99 % av de riktige taggene. *Presisjonen* er på 95,4 %, noe som betyr at 95,4 % av de taggene som blir stående igjen, er riktige. På nynorsksiden ligger *recall* og *presisjon* på henholdsvis 98,7% og 93,6%.

2.3 Speilmetoden

2.3.1 Innledning

I det siste tiåret har det oppstått en økende interesse for å bruke parallellkorpora som en kilde til informasjon om ordbetydninger (Dyvik, 1998/2002; Ide, 1999; Resnik & Yarowsky, 2000; Diab & Resnik, 2000). Parallellkorpora som informasjonskilde for ordbetydninger er interessant fordi kontrastivt flertydige ord ikke forventes å ha oversettelser med samme flertydighet i et annet språk, og man venter heller ikke at flere ord i samme språk skal ha den samme kontrastive flertydighet. Oversettelsen av et flertydig ord til et annet språk kan dermed fungere som en diskriminator mellom ulike betydninger av ordet. I eksemplene (1-2) under oversettes f.eks. det flertydige norske substantivet *tak* til engelsk som *roof* i bygningstak-betydningen, og som *hold* i gripetak-betydningen:

(1) *Far hjalp naboen vår å bygge **tak**.*
*Father helped our neighbour building a **roof**.*

[9] Tilgjengelig fra nettsiden: http://www.hf.uio.no/iba/prosjekt/.
[10] Informasjon om taggerens prestasjonsnivå foreligger fra nettsiden: http://www.tekstlab.uio.no/norsk/bokmaal/ og ../norsk/nynorsk/.

(2) *Først må du ta godt **tak** i kassen.*
*First you have to get a firm **hold** of the box.*

Å bruke parallellkorpus som ressurs for orddisambiguering er dessuten teoretisk interessant med henblikk på det generelle spørsmålet om hvilke kriterier som bør, eller kan, legges til grunn for å trekke skiller mellom ordbetydninger. Håpet er at betydningsskiller realisert på tvers av språk kanskje kan fungere som et konsekvent kriterium for å definere betydningsskiller innenfor ett språks inventar av ord og uttrykk. (I hvilket fall målet ikke nødvendigvis er "perfekte" skiller som er i samsvar med intuisjonene til enhver person, snarere kan det kanskje være realistisk å bruke informasjon fra parallellkorpora for å finne fram til et håndfast, rimelig og konsekvent kriterium som kan automatiseres.)

Forskningsprosjektet "Fra parallellkorpus til ordnett" (2001-2004) er tilknyttet Universitetet i Bergen og HIT-senteret, og er finansiert av Meltzerfondet og Norges forskningsråd. Målet i dette prosjektet er å videreutvikle og teste Dyviks speilmetode som automatisk avleder semantiske ordnett (Dyvik, 1998/2002). Innenfor prosjektet arbeides det også med å tilrettelegge ENPC-korpuset for å kunne applisere speilmetoden direkte på korpuset (det vil si å utarbeide en ordparallellstilling av korpuset). I tillegg kommer arbeid med å lage en manuelt basert "gullstandard" som et slikt automatisk ekserpert materialet kan evalueres mot, samt evaluering underveis av resultatene. Et vellykket resultat vil innebære at deler av arbeidet med å utvikle et norsk ordnett som en ressurs for norsk språkteknologi vil kunne automatiseres.

Enkelt sagt kan metoden sies å gå ut på å "speile" leksikalske enheter frem og tilbake oversettelsesmessig mellom to språk, derav navnet "speilmetoden". En *leksikalsk enhet* defineres i det følgende som enkeltord (nærmere bestemt lemmaer) og flerordsuttrykk som etter visse kriterier kan fastslås å ha en utskillbar korrespondent i de parallelle setningene.

Den grunnleggende hypotesen bak speilmetoden (Dyvik, 2002) at hvis to leksikalske enheter *x* og *y* i et gitt språk L1 er semantisk nærbeslektede, så vil deres respektive sett av mulige oversettelser i et språk L2 sannsynligvis overlappe. En videre antagelse er at ord med en generell betydning vil ha flere mulige oversettelser enn ord med mer spesifikke betydninger. Derav følger at om *x* er et hyponym (underbegrep) til *y*, så burde de mulige oversettelsene av *x* være et subsett av de mulige oversettelsene av *y*. Om to leksikalske enheter *x* og *y* derimot ikke har noe semantisk slektskap, forventer vi at de maksimalt skal ha én oversettelsespartner felles som da er kontrastivt flertydig (For eksempel slik de semantisk urelaterte engelske substantivene *dish* og *claim* begge har norsk *rett* som felles oversettelsespartner). Uten mer enn én slik felles oversettelseskorrespondent faller de leksikalske enhetene *x* og *y* inn under ulike semantiske felt, hvilket med andre ord betyr at de skilles ut som ulike betydninger.

Speilmetoden tar utgangspunkt i oversettelseskorrespondanser i parallellkorpuset ENPC, i form av at leksikalske enheter i hvert språk registreres med settene av sine mulige oversettelser i det motstående språket. Disse settene av oversettelseskorrespondanser brukes som grunnlag for å utføre følgende operasjoner (Dyvik, 2002):

1. Kalkulere hver leksikalske enhets ulike betydninger relativt til det gitte motstående språket. Dette impliserer med andre ord å kalkulere graden av flertydighet for en leksikalsk enhet i forhold til det motstående språk.
2. Kalkulere hvilke ordbetydninger innenfor et språk som tilhører samme semantiske felt, i.e. som er semantisk relaterte.
3. Tildele alle ordbetydninger innen et semantisk felt et sett av semantiske trekk. Hver ordbetydning er opphav til minst et unikt trekk, og arver dessuten trekk fra eventuelle ordbetydninger som ifølge data fra parallellkorpuset er overbegreper. På denne måten blir en ordbetydnings relasjon til andre ordbetydninger i det semantiske feltet implisitt kodet inn for hver ordbetydning.
4. Derivere et semantisk ordnett med ordbetydninger som noder: På grunnlag av overlapping mellom ordbetydningenes respektive sett av trekk, systematiseres ordbetydningene i en semi-lattice som visualiserer de semantiske relasjonene som ifølge parallellkorpuset holder mellom dem (f.eks. med hyponymi-/hyperonymi-relasjoner og nærrelasjoner).
5. Disse semantiske relasjonene kan så brukes for å generere tesaurus-oppslag for hver leksikalske enhet.

Med hensyn til forsøket på å lage en automatisk betydningstagger er kun informasjonen fra speilmetodens første operasjon direkte relevant, d.v.s. steget som utleder en leksikalsk enhets sett av betydningspartisjoner.

Fordi ekserperingen av oversettelseskorrespondansene som gir betydningspartisjoner foreløpig må foregå manuelt, var det i første omgang aktuelt å velge ut ett rimelig frekvent flertydig ord fra korpuset. Valget falt på ordformen *rett*, som gir oss de to lemmaene *rett*N og *rett*A. De to lemmaene ble holdt atskilt under ekserperingen, hvilket også resulterer i atskilte sett av betydningsskiller for dem.

I del (2.3.1) under følger en presentasjon av prinsippene som ble definert som retningslinjer under det manuelle ekserperingsarbeidet.

2.3.2 Ekserperingsprinsipper

Hvis det skal ha noe for seg å bruke parallellkorpus som en rikest mulig informasjonskilde om leksikalske enheters betydninger (og med minst mulig feilinformasjon), kan vi kort si der er to hovedhensyn som motiverer formuleringen av prinsippene: For det første ønsker vi å motivere at registrering ikke skal skje i tilfeller hvor det intuitivt synes klart at der ikke foreligger en god korrespondent til søkeuttrykket i den parallellstilte setningen. På den andre siden ønsker vi heller ikke å være for strenge, siden det er ønskelig å hente ut mest mulig informasjon fra korpuset.

Ekserperingsprinsippene presentert her er utarbeidet i samarbeid med de øvrige som har vært involvert i ekserperingsarbeid innenfor forskningsprosjektet "Fra Parallellkorpus til Ordnett", og er i stor grad basert på Thunes (2003). Som Thunes påpeker, er det vanskelig å definere prinsipper som kan gi svar i ethvert tvilstilfelle som måtte oppstå under ekserpering, og det er uunngåelig at ekserperingsarbeidet får heuristiske innslag og involverer skjønnsmessige avgjørelser. De følgende definisjoner er resultatet av et forsøk på å lage prinsipper som er presise nok til å la seg anvende, men samtidig ligger nær opp til det man intuitivt mener er riktig. I det følgende fokuseres det på søkeuttrykk som uttrykker argumenter

(f.eks. substantiver) eller ikke-argumenter (f.eks. adjektiver). For en mer omfattende diskusjon av ekserperingsprinsipper benyttet i forskningsprosjektet, refereres det til Thunes (2003).

Utarbeidingen av ekserperingsprinsippene i forbindelse med denne oppgaven må ses i lys av hva som er relevante data for å skulle benytte speilmetodens output i et program som tagger ordformer i et korpus semantisk.

I. Typen søkeuttrykk som tilordnes oversettelseskorrespondanser

Vi er interessert i å bruke speilmetoden for å hente ut informasjon om leksikalske enheters betydning(er). Siden alle ordformer innen et bøyningsparadigme som regel deler semantiske egenskaper, ville det resultere i unødig fattige data ved å registrere oversettelseskorrespondanser for hver av bøyningsformene separat. I stedet registreres oversettelseskorrespondansene for alle ordformer innen et bøyningsparadigme i fellesskap, det vil si vi registrerer oversettelseskorrespondanser for lemmaer.

I enkelte tilfeller er oppslagsformen identisk for flere lemmaer, slik for eksempel adjektivet og substantivet *rett* har samme lemma. Prinsipielt krever ikke speilmetoden informasjon om ordklasser for å skille ut de ulike betydninger av en ordform som kan tilhøre ulike ordklasser. Siden ENPC-korpuset er lemmatisert og tagget for pos synes det likevel overflødig å slå sammen ordkategorier i registreringen, gitt at flertydighet på tvers av ordklasser enklere kan løses ved å benytte pos-taggene. Følgelig står vi fast ved valget om å registrere oversettelseskorrespondanser for søkeuttrykk basert på lemmaer, definert som samlingen av alle ordformer som tilhører samme bøyningsparadigme.

Det forekommer av og til at et søkeuttrykk ikke kan sies å korrespondere med ett utskillbart lemma i det motstående språket. I enkelte tilfeller registreres derfor et flerordsuttrykk som korrespondent til søkeuttrykket, hvilket innebærer at det i neste omgang er dette flerordsuttrykket som blir gjenstand for søk. Dermed vil begrepet leksikalsk enhet/et søkeuttrykk omfatte både enkeltstående lemmaer, og flerordsuttrykk som er redusert til lemmaer i den grad det er forenlig med søkeuttrykkets identitet.

II. "Oversettelsesmessig korrespondanse" mellom søkeuttrykk og en leksikalsk enhet i det motstående språket

Enkelt kan man si at hvis omgivelsene *rundt* søkeuttrykket semantisk korresponderer med omgivelsene rundt en leksikalsk enhet i det motstående språket, så registrerer vi en korrespondanse.

Mer formelt kan det uttrykkes slik (Thunes, 2003): Gitt at et søkeuttrykk *a* er inneholdt i kildesetningen K i et språk L1, og gitt at K er parallellstilt med målsetningen M i et språk L2. Om søkeuttrykket for eksempel inngår i en leddsetning, så er det denne leddsetningen vi regner som K, og M er den setningsenheten i motstående språk som tilsvarer K. Vi sier at *a* korresponderer med et uttrykk *b* i målsetningen M i L2 hvis følgende betingelser er tilfredsstilt:

i: Generell regel for ekserpering:

M og K må ha tilstrekkelig like argumentstrukturer til at uttrykkene i *a*s omgivelser står i de samme semantiske relasjonene til hverandre og til *a* som de korresponderende uttrykkene i *b*s omgivelser gjør til hverandre og til *b*. Det er ikke et krav at lenkningen mellom syntaktiske konstituenter og semantiske roller er identisk for de korresponderende relasjonene og tilhørende argumenter, og det må ikke nødvendigvis være likt antall *syntaktiske* ledd omkring henholdsvis *a* og *b*. Omgivelsene defineres som "tilstrekkelig like" etter følgende kriterier:

ii: Når *a* og *b* uttrykker argumenter (eksemplifisert bl.a. ved *rett*N):

Hvis *a* og *b* uttrykker argumenter, må disse være tildelt samme type semantiske rolle. Videre, hvis *a* og *b* er referensielle uttrykk, må de være koreferente, altså forankret i samme referent i diskursen. Eksemplene (3) og (4), hentet fra ENPC[11], illustrerer henholdsvis hva vi ifølge kriteriene vil og ikke vil registrere. I (3) ser vi at betingelsene er oppfylt for å ekserpere *order* som en korrespondent til søkeuttrykket *rett*: K identifiseres som leddsetningen [*Når en rett var klar til servering*] og M identifiseres som [*When an order was ready*]. De korresponderende NPene *rett* og *order* er begge tildelt patiensrollen. Gitt at det ikke er et krav at antallet syntaktiske konstituenter er identisk mellom K og M, er det uproblematisk at den opsjonelle PPen *til servering* ikke har noen uttrykt korrespondent i den engelske setningen.

(3) [$_K$ *Når en* ***rett*** *var klar til servering*], *ringte madame med en liten klokke på kjøkkenet, og hennes mann hevet øyebrynene i påtatt irritasjon.*
(PM1T)
[$_M$ *When an order was ready*], *Madame would clang a bell in the kitchen and her husband would raise his eyebrows in pretended irritation.*

I eksempelet (4) under er kriteriene derimot ikke oppfylt for å registrere en korrespondent. Ut fra omgivelsene kan vi se at den potensielle kandidaten som korrespondent til *rett* ville være *entitled*. Siden *ha* og *be* tildeler ulike semantiske roller til sine respektive omgivelser, er imidlertid ikke betingelsene våre oppfylt. Dette formelle kriteriet er også i samsvar med den umiddelbare intuisjonen (hvilket for ordens skyld *ikke* er et kriterium!) om at å "ha retten til noe" (den norske setningen) ikke er helt det samme som å "være berettiget til noe" (som er hva den engelske setningen uttrykker). Riktignok kunne vi si at sekvensen *ha rett til* som leksikalsk enhet korresponderer med sekvensen *be entitled to*, men *rett* korresponderer ikke alene med *entitled*.

(4) [$_K$ *Arbeidstakerne og/eller deres representanter har* ***rett*** *til å klage*], *i samsvar med (...)*
(EEA1T)
[$_M$ *Workers and/or their representatives are entitled to appeal*], *in accordance with (...)*

[11] Eksempler hentet fra ENPC er oppført med en kode i parentes mellom de korresponderende setningene, som refererer til teksten i ENPC som den første oppførte setningen er hentet fra. En oversikt over hvilke tekster de forskjellige kodene peker til, finnes på http://www.hf.uio.no/iba/prosjekt/>. Hvis et eksempel ikke er ført opp med en slik kode, er det et konstruert eksempel.

iii: Når *a* og *b* uttrykker ikke-argumenter (eksemplifisert bl.a. ved *rett*A):

Hvis *a* og *b* uttrykker ikke-argumenter, som f.eks. adverbielle ledd, må også disse være koreferente, i den forstand at de refererer til samme situasjon, lokasjon eller egenskap og modifiserer korresponderende enheter, f.eks. verb (Thunes, 2003). Det er ikke nødvendig at *a* og *b* er nærrelaterte uttrykk. Ikke-argumenter kan være særtilfeller av relasjoner: Et adverb som modifiserer noe, kan sees som en en-plass relasjon som tar den modifiserte enheten som argument. Korrespondanse mellom ikke-argumenter kan illustreres med eksempel (5), hvor adverbet *rett* er søkeuttrykk.

(5) *Det vil være mange som blomstrer i dette finansmiljøet,* [$_K$ *og atskillig flere som kan gå* ***rett*** *ad undas.*]
(JH1T)
There will be many who flourish in this environment of finance,[$_M$ *and a great many more who can go straight to hell.*]

I den engelske setningen finner vi adverbet *straight*, og betingelsene for å si at *rett* korresponderer med *straight* er oppfylt på følgende måte: *rett* uttrykker en en-plass relasjon som tar som argument den relasjonen som er uttrykt av adverbialet *ad undas*, mens *straight* uttrykker en en-plass relasjon som tar som argument den relasjonen som er uttrykt av adverbialet *to hell*. De enhetene som er argumenter for henholdsvis *rett* og *straight*, korresponderer på selvstendig grunnlag.

III. Særtilfeller hvor vi går bort fra hovedreglene (I og II) for ekserpering

i: Formlikhet og avledningsforhold mellom adjektiv og adverb

Både i norsk og engelsk forekommer det at en og samme ordform kan opptre både som adjektiv og som adverb. I eksemplene (6) og (7) under fungerer *rett* som adjektiv i (6) og som adverb i (7):

(6) *Heldigvis at jeg plasserte den rosakledte konfirmasjonsdukken fra telegramtablået på plass i* ***rett*** *tid.*
(CL1)
Fortunately, I put the pink-clad Confirmation doll from the telegram tableau in her place at the right time.

(7) *"Hvis jeg har forstått Simon* ***rett****, er det (...)".*
(RDA1T)
"If I understand Simon rightly, it 's (...)".

Som Thunes (2003) påpeker, er der videre både i norsk og engelsk grupper av adjektiver og adverb som har en felles leksikalsk rot, og der adverbene er morfologisk avledet fra adjektivene:

> *I norsk gjelder dette adverb som er identiske med nøytrumsformen av tilsvarende adjektiv, f.eks. adjektivet tørr og det avledede adverbet tørt. I engelsk gjelder dette særlig adverb som er dannet fra adjektiver ved tillegg av suffikset -ly, og det gjelder også danning av adverb ved de mindre frekvente suffiksene -ward, -ways og -wise (kfr. Miller 1998: 60).*
>
> *(Thunes, 2003)*

Forekomstene av henholdsvis *right* og *rightly* i eksemplene (6) og (7) illustrerer dette. Som illustrert i (8), forekommer det dessuten også at en ordform kan fungere både adjektivisk og adverbialt, selv om der finnes en mulig avledningsform. (8) viser at ordformen *deep* kan fungere adverbialt i setningen, selv om vi vet at der eksisterer en adverbial avledningsform *deeply*.

(8) ***Deep*** *within the hardness and firmness of this man lay a soft, lonely place.*
(BV1T)
Rett innenfor hardheten og fastheten i denne mannen lå et mykt ensomt sted.

Slike tilfeller reiste tvil om hvordan man skal behandle adjektiver og adverb i ekserperingen av oversettelseskorrespondanser. I henhold til prinsippet om at vi søker på lemmaer burde det skilles mellom adjektiver og adverb i ekserperingen. Under ekserperingen oppstod det imidlertid ofte tvil om hvordan man i praksis skal behandle tilfeller som *deep* ovenfor, hvor én og samme ordform kan sies å ha adverbial *funksjon* selv om vi vet at der eksisterer en adverbial avledningsform.

Disse fenomenene; formlikhet og avledningsforhold mellom adjektiver og adverb, samt syntaktisk funksjon i en gitt kontekst, har ledet til å gjøre et unntak fra avgjørelsen om å skille mellom ordklasser i ekserperingen (Thunes, 2003): Når adjektiver og adverb står i avledningsforhold til hverandre, som f.eks. *tørr* - *tørt*, og *right* - *rightly*, og hvor medlemmene i et slikt par synkront sett har samme leksikalske rot, velger vi å oppheve kategoriskillet mellom adjektivet og adverbet, med hensyn til både søkeordet og dets korrespondenter.

Dette innbærer at hvis søkeuttrykket var *rett* som i (6-7) ovenfor, så søkes det etter både adjektiviske og adverbiale forekomster av søkeuttrykket, og det registreres motparter for begge typer forekomster under ett. Det registreres heller ikke noe kategoriskille blant de korrespondentene som finnes i den parallelle teksten. Ut fra eksemplene i (6-7) ovenfor ville vi da registrere ganske enkelt *right* som korrespondent til *rett*.

Imidlertid, hvis søkeuttrykket er *hardly*, må dette sies å ha en annen leksikalsk rot enn både adjektivet *hard* og adverbet *hard*, og vi må da ta hensyn til dette skillet. Dette innbefatter også enkelte flerordsuttrykk, hvor det vil være mest naturlig å registrere den leksikalske enheten slik den forekom i teksten. I (9) under ville det være unaturlig å registrere *just_bare* som en korrespondent til søkeuttrykket *såvidt*, og istedet registreres korrespondenten som *just_barely*:

(9) *(...), han kan se en rosa ransel som* ***såvidt*** *stikker fram.*
(LSC1)
He can see a pink school bag just barely sticking out.

ii: Korrespondanser i form av særnavn eller et medlem av en lukket ordklasse

Under ekserperingen av *rett* syntes det i enkelte tilfeller velmotivert å la være å registrere en korrespondent, selv om omgivelsene for henholdsvis *a* og *b* tilfredsstiller kriteriene angitt i (II) ovenfor. Nærmere bestemt dreier dette seg om tilfeller hvor et substantivisk søkeuttrykk korresponderer med et særnavn, og enkelte tilfeller hvor et søkeuttrykk korresponderer med et medlem fra en lukket ordklasse.

I utgangspunktet er det ikke en forutsetning at argumentene uttrykt av *a* og *b* er nærrelaterte; de kan være presiseringer eller depresiseringer av hverandre (Thunes, 2003). I eksempelet under korresponderer for eksempel søkeuttrykket *middag* med engelsk *food*, og intuitivt ville vi si det foreligger en hyponymirelasjon mellom *middag* og det mer generelle *food*. (At middag er en mulig type måltid/mat, men mat inntas ikke nødvendigvis i form av en middag.)

(10) *Siden serverer hun ham **middag** i taushet.*
(ROB1T)
Then she serves him food in silence.

Generelt ønsker vi å registrere en slik korrespondanse mellom to leksikalske enheter der den ene har enten et mye snevrere eller mye videre sett av denotata enn den andre (som i eksempel (10) ovenfor), selv om de ikke nødvendigvis vanligvis finnes i tospråklige ordbøker. Korrespondanser av denne type gir informasjon om hypo- og hyperonymirelasjoner mellom leksikalske enheter, og denne typen informasjon ønsker vi å få representert i ordnettet. Siden ekserperingens mål er å få informasjon om leksikalske enheters betydninger, er det likevel ønskelig å luke ut oversettelseskorrespondanser som må sies å være spesifikke for *forekomsten* av et søkeuttrykk (og dens korrespondent), og som ikke har med søkeuttrykkets *type* å gjøre.

Dette innebærer for det første at særnavn (NPer med fast referanse) ikke inkluderes i ekserperingsmaterialet. Særnavn kan vanskelig sies å egentlig ha noen annen "betydning" enn at de hjelper oss å referere til en presis entitet i verden, og de er derfor ikke interessante som informasjonskilde i ordnettet. I eksempelet under er søkeordet *medalje*, og den mulige korrespondenten i engelsk er nomenet *DSC*. Siden *DSC* er et særnavn, registreres den ikke som korrespondent selv om omgivelsene ellers tilfredsstiller kriteriene for å ekserpere.

(11) *Etter at Gary fikk sin **medalje** gjort om til et askebeger av Buddy Torgeson , (...)*
(SK1T)
After Gary had his DSC turned into an ashtray by Buddy Torgeson, (...)

En videre modifikasjon av hovedregelen gjelder de tilfeller hvor søkeuttrykkets korrespondent tilhører en lukket ordklasse. Som eksempel kan det forekomme at et substantivisk søkeuttrykk opptrer i en referensiell nominalfrase (f.eks. *medalje*) som korresponderer med en anafor (f.eks. *it*) i den parallelle teksten[12]. Når grammatikaliserte

[12] I samsvar med Thunes (2003) skal begrepet anafori her bety ulike typer uttrykk som refererer tilbake til en entitet som allerede er introdusert i diskursen av et annet refererende uttrykk. Pronomina (*han, min*) og demonstrativer (*denne*) kan kalles **grammatikaliserte anaforer**, mens medlemmer av kategoriene kvantorer (*mange*) og enkelte adjektiver (*slik*), har mulighet for å opptre med anaforisk funksjon i bestemte kontekster. Substantiver, som utgjør en åpen klasse, er i utgangspunktet ikke anaforer, men kan opptre anaforisk når det

anaforer som pronomina og demonstrativer alene opptrer som korrespondenter for (leksikalske) søkeord, er det ikke ønskelig å registrere korrespondansene - Altså ville vi ikke f.eks. registrere pronomenet *it* som korrespondent for søkeordet *medalje*, selv om betingelsene for oversettelsesmessig korrespondanse ellers skulle være oppfylt.

Bakgrunnen for dette valget er at vi kun er interessert i å registrere korrespondanser på *typenivå*. Vi kan for eksempel tenke oss konsekvensene det ville få å inkludere grammatikaliserte anafora med tanke på å benytte speilmetodens betydningsskiller til et praktisk formål som automatisk betydningstagging: Hvis en grammatikalisert anafor som *it* tilfeldigvis ble registrert som en korrespondent til *rett*N, så ville anaforen innordnes en bestemt betydningspartisjon *x* for *rett*N. Dette ville i seg selv vanskelig kunne forsvares prinsipielt, siden grammatikaliserte anafora ikke i seg selv har en referensiell betydning, men får sin betydning fra et tidligere introdusert ord. Med tanke på et praktisk formål som automatisk betydningstagging, ville det dessuten være lite ønskelig at en tilfeldig anafor skulle kunne "avgjøre" at en gitt forekomst av *rett*N er å kategorisere som betydning *x*.

Hvis man derimot har et tilfelle hvor en grammatikalisert anafor (pronomina eller determinativer) inngår som en del av uttrykket *b*, kan det likevel være aktuelt å registrere korrespondenten. Det kan da bli nødvendig å vurdere skjønnsmessig om en korrespondent skal ekserperes eller ikke.

Under arbeidet med ekserpering viste det seg imidlertid å være alt for enkelt å si at når søkeuttrykket tilhører en åpen ordklasse, så skal vi ikke registrere en korrespondent fra en lukket ordklasse. Der er typer av lukkede klasser som kan synes motivert å ta med i nettverket; en av dem er kvantorer som *mange*, *alle*, *noen*, *få*, *ingen*. Som eksempel ble *rett*A registrert å korrespondere med engelsk *rather*. Under søk på korrespondenter for dette ordet, ble *rather* funnet å korrespondere med adjektivet *mye*, som i sin tur korresponderer med kvantoren *many*:

(12) *(...) og Platon kunne ikke ha kjent i så **mye** som ti år den mannen som skulle inngi ham en livsvarig kjærlighet til tanken, og som (...).*
(JH1T)
(...), and Plato could not have known for as many as ten years the man who was to inspire him with a lifelong devotion to thought and whose (...).

I den videre ekserperingen av *many* søkes det da på korrespondenter tilknyttet denne ordklassen, dvs vi søker etter alle korrespondenter for kvantoren *many*.

Siden ekserperingsmaterialet for denne oppgaven ikke er svært omfattende, forekom det tross alt sjelden at man støtte på tilfeller hvor det syntes aktuelt å registrere korrespondenter fra lukkede ordklasser – Følgelig gir heller ikke speilmetodens resultater i denne oppgaven noe tydelig bilde av konsekvenser det eventuelt får å registrere ord fra lukkede ordklasser. Men med tanke på en mer omfattende automatisk ekserpering når ENPC er parallellstilt på ordnivå, ville det være interessant å se på hvilke konsekvenser det har om ord fra lukkede i større utstrekning inngår i ordnettet sammen med ord fra åpne ordklasser.

Ekserperingsprinsippene vil diskuteres i (2.5), hvor vi diskuterer speilmetodens resultater for lemmaene i denne oppgaven. Som vi vil se, later ekserperingsprinsippene til å ha fungert godt med hensyn til speilmetoden.

---

foreligger synonymi-, hypo- eller hyperonymirelasjoner mellom dem - F.eks. kan den referensielle NPen *maten* sies å opptre anaforisk i forhold til NPen *middagen*.

2.3.3 Speilmetodens utarbeiding av betydningsskiller

Fordi det i evalueringen i neste delkapittel er hensiktsmessig å referere til de ulike ekserperingsstegene som ble foretatt for å få ut betydningsskiller, vil vi i dette delkapittelet bruke ekserperingsmateriale for *rett*N til å illustrere de fire "speilings"-stegene som er nødvendig for å få ut en leksikalsk enhets sett av betydningspartisjoner.

Første steg er å hente ut settet av oversettelsesrelasjoner mellom *x*, i vårt eksempel *rett*N, i L1 og leksikalske enheter i L2. Med Dyviks terminologi kan dette kalles *first translational image* of *x*, som vi kortere kan utrykke som *first t-image* (*x*). I henhold til de data som foreligger i ENPC ble første oversettelsesbilde for *rett*N som følger:

(13) **First t-image (rettN)**
(claim court course dish entitlement food justification law option order rightN specialN supper)

Dette settet inneholder alle de leksikalske enheter i L2, i vårt tilfelle engelsk, som ble funnet å korrespondere med søkeuttrykket *rett*N i henhold til ekserperingsprinsippene. *First t-image* (*rett*N) kan beskrives som et uordnet sett, hvor vi intuitivt kan se at minst to betydningsskiller for *rett*N utkrystalliserer seg: Vi aner at ord som *claim* og *entitlement* "hører sammen", mens ord som *dish* og *food* kan tilordnes en annen betydning. For å få ordene sortert i slike betydningspartisjoner (*sense-partitions*) maskinelt behøves imidlertid ytterligere informasjon om oversettelsesmessige overlappinger mellom ord i L1 og L2, hvilket bringer oss til steg to.

Steg to er å finne oversettelseskorrespondansene for hvert respektive ord inneholdt i *first t-image* (*rett*N). Dette utgjør det inverse oversettelsesbildet til *rett*N, uttrykt som *inverse t-image* (*rett*N):

(14) **Inverse t-image (rettN)**

(rightN	(adgang konsesjon krav rettighet rett))
(court	(domstol rett hoff slott plass gård gårdsplass palass sak fylkesmann ting domstolsbehandling))
(entitlement	(rett rettighet adgang))
(law	(lov rett jus regelverk juridisk))
(justification	(argument begrunnelse berettigelse rettferdiggjøring rett))
(dish	(tallerken gryte skål rett fat oppvask servise måltid kopp kar))
(supper	(aftensmat aftens aftensbord middag mat supé rett kvelds kveldsmat))
(food	(kost kosthold føde næring mat matvare matvei middag mattilskudd jordbruksprodukt rett))
(course	(kurs rett undervisning))
(special	(rett lunsj))
(option	(mulighet vei_å_gå tilbud rett løsning valgmulighet))
(order	(orden ordre måte bestilling rekkefølge medalje stand klasse sjikt kommando signal befaling påbud pålegg lov vedtak bestemmelse kjennelse forelegg regel rett system))
(claim	(rett krav fordring påstand))

Som vi ser, inneholder disse settene til en viss grad ord som kan sies å være relatert til det opprinnelige ordet *rett*N (for eksempel *krav*, *rettighet* i første sett). I tillegg forekommer der også ord som ikke er semantisk relatert til *rett*. Disse kommer med som en følge av at de engelske leksikalske enhetene som ekserperes selv kan være flertydige. (For eksempel ord som *slott, gård* og *gårdsplass* i det andre settet i (14), som er korrespondenter til engelsk *court*.)

Tredje steg i speilmetoden er å ta unionen av alle settene som utgjør det inverse oversettelsesebildet av *rett*N med hensyn til L2, og å ta bort det opprinnelige ordet *rett*N[13]. Hva man da får er et samlet sett av L1-ord (i vårt tilfelle, norske ord), og dette settet utgjør grunnlaget for å ekserpere *rett*Ns andre oversettelsesbilde; *second t-image* (*rett*N).

Det ekserperte s*econd t-image* (*rett*N) består følgelig av et stort sett av subsett; et subsett for hvert av de norske ordene fra *inverse t-image* (*rett*N). (Dette beløper seg til 97 subsett.) Siden ord i både det første og inverse oversettelsesbildet av *rett*N selv kan være flertydige, vil *rett*Ns andre oversettelsesbilde inneholde mange ord som ikke har noe å gjøre med betydningen(e) til *rett*N. Formålet med å ekserpere *rett*Ns *second t-image* er at ord struktureres i subsett: Ved å begrense *second t-image* (*rett*N) til kun de ord som forekom i *first t-image* (*rett*N), oppnår vi at den uordnete listen av ord i *rett*Ns første oversettelsesbilde nå forekommer innenfor strukturerte subsett. Dette begrensede oversettelsesbildet kan vi kalle *rett*Ns *restricted second t-image*:

(15) **Restricted second t-image(rettN)**
((order rightN claim)
(order rightN law)
(justification court)
(order option)
(order law)
(rightN entitlement)
(supper specialN)
(supper food)
(supper dish)
(claim)
(course)
(court)
(dish)
(food)
(justification)
(law)
(option)
(order)
(rightN)
(supper))

Som vi ser, overlapper enkelte av subsettene, og det er dette som utgjør basisen for å finne betydningspartisjonene for et en gitt leksikalsk enhet. Ved å slå sammen de subsettene som overlapper ved minst ett ord, står vi til slutt igjen med partisjonering av alle ordene fra *restricted t-image*(*rett*N). Partisjoneringen av det første oversettelsesbildet av *rett*N gir fire sett, som vil refereres til som *rett*Ns sett av betydningspartisjoner, eller betydningsskiller:

[13] Oversettelseskorrespondansene for alle ord i unionen av *Inverse t-image* (*rett*N) utgjør grunnlaget for å finne det andre oversettelsesbildet av *rett*. Grunnen til at ordet *rett* da må ekskluderes er at målet er å få ut betydningsskillene for nettopp *rett*N. Fordi betydningsskillene er basert på overlapping i *rett*s andre oversettelsesbilde, er selve ordet *rett* nødt til å ekskluderes: Ellers ville alle sett i andre oversettelsesbilde direkte eller indirekte hatt overlappingsrelasjoner.

(16) **Sense partitions of first t-image of "rettN":**
((course)
(court justification)
(claim entitlement law option order rightN)
(dish food specialN supper))

For denne oppgavens formål er det disse betydningspartisjonene som er relevant: I metoden for automatisk betydningstagging er det et lemmas betyningspartisjoner, som i (16), som legges til grunn for å tagge forekomster av *rett*N i ENPC for betydning.

Betydningsskillene illustrert i (16) ovenfor kan også presenteres i form av et mer informativt oppslag etter mønster av tradisjonelle tesaurusoppslag (punkt (5.) s. 21), hvor eventuell informasjon om underbetydninger, synonymer/relaterte ord og hyponymi/hyperonymi-relasjoner for hver betydning er inkludert (Dyvik, 2002). I evalueringen av *rett*N og *rett*A i følgende delkapittel vil vi benytte slike tesaurusoppslag, siden informasjonen om semantiske relasjoner tilknyttet hver ordbetyding litt bedre "illustrerer" hvert betydningsskille. Det må da understrekes at evalueringen ikke vil fokusere sterkt på oppslagets tilleggsinformasjon ut over selve betydningsskillene, siden speilmetodens formål i denne oppgaven begrenser seg til å supplere oversettelsesbaserte betydningsskiller for et flertydig ord.

Speilmetoden er fremdeles under videreutvikling innenfor prosjektet "Fra parallellkorpus til Ordnett", deriblant med hensyn til kriterier for å generere informasjon om semantiske relasjoner for hver ordbetydning i et tesaurusoppslag. I tilknytning til denne hovedoppgaven har det blitt gitt tilgang på tesaurusoppslag hvis bakenforliggende definisjoner er oppdaterte i forhold til hva som er beskrevet i Dyvik (2002). Siden definisjonene til en viss grad avviker fra litteraturen det vises til. synes det derfor på sin plass å spesifisere for ordens skyld hvordan denne oppgavens presenterte tesaurusoppslag er generert.

Tesaurusoppslagets hovedbetydninger (*sense 1, sense 2*, etc) er gitt av speilmetodens betydningspartisjoner (som i (16)), og det er dermed disse hovedbetydningene evalueringen fokuserer på. Øvrig informasjon om en betydnings eventuelle inndeling i underbetydninger, samt informasjon om synonymer, relaterte ord og hyponymi-/hyperonymirelasjoner, er basert på speilmetodens kalkulering av semantiske trekk innenfor hver ordbetydnings semantiske felt (et steg som følger *etter* at betydningspartisjonene er fastlagt, jf. punkt (3.) s. 21). Speilmetoden innordner hver betydning i et semantisk felt, hvor ordbetydningen er hierarkisk plassert i forhold til andre semantisk relaterte ord. Hver ordbetydning er opphav til minst ett trekk i sitt semantiske felt; ordbetydningens *opphavstrekk*. Siden underbegreper arver trekk fra sine overbegreper i det semantiske feltet, kan en ordbetydning også være assosiert med *arvede* trekk. Som eksempel viser (17) under de semantiske trekk som genereres for betydning 4 av *rett*N (i det følgende referert til som *rett*N_4). Vi ser at ordbetydningen er opphav til det siste trekket (som vi ser ved at ordet finnes i trekknavnet). De øvrige trekkene er arvet fra de ordbetydninger som dominerer *rett*N_4, og som ifølge speilmetodens ordnett dermed utgjør en form for overbegreper.

(17) **Semantiske trekk assosiert med *rett*N_4:**
[mat1|supper2]
[middag1|food5]
[måltid1|dish3]
[fat1|dish3]
[kar1|dish3]
[rettN4|specialN1]

Det er verdt å merke seg at mens selve ordbetydningene oppført i et tesaurusoppslag er fastlagte av betydningspartisjonene, genereres tesaurusoppslagets trekkbaserte informasjon om semantiske relasjoner for hver ordbetydning ut fra parameterverdier som kan endres av brukeren i dannelsen av et oppslag. Informasjon ut over selve opplistingen av ordbetydninger er altså av mer variabel art.

Parameteren *OverlapThreshold* avgjør inndelingen av underbetydninger (*subsenses*) for et teasurusoppslags betydninger. Underbetydninger er ment å reflektere eventuelle ulike aspekter av ordbetydningen som en ordbetydning, for eksempel *rett*N_4, denoterer. Underbetydninger defineres ved å måle i hvilken grad la oss si trekket [måltid1|dish3] i (17) overlapper med [middag1|dish3] i hele det semantiske feltet. Antagelsen er som følger: Hvis disse to trekkene fanger inn samme aspekt (underbetydning) av betydningen som denoteres av *rett*N_4, så burde disse to trekkene ofte forekomme sammen også for andre ordbetydninger enn *rett*N_4 i det semantiske feltet. På grunnlag av et mål på hvor ofte trekk overlapper i hele det semantiske feltet kan en hovedbetydning altså inndeles i underbetydninger, og antall underbetydninger avhenger av hvilken verdi brukeren velger å sette for parameteren O*verlapThreshold*. I evalueringen av *rett*N og *rett*A anvendes speilmetode-algoritmens "default"-verdi for denne parameteren, 0.05 (Dyvik, 2002).

En parameter *SynsetLimit* avgjør genereringen av en ordbetydnings synonymer, relaterte ord (som antas å stå noe fjernere fra den aktuelle ordbetydningen enn synonymer) og hyperonymer/hyponymer. Beregningen av semantiske relasjoner for en gitt ordbetydning *x* er basert på hvor mange ordbetydninger i hele det semantiske feltet som er tilordnet hvert av *x* sine semantiske trekk. Ordbetydningene som er assosiert med ett og samme semantiske trekk kan refereres til som trekkets *denotasjon*. En ordbetydning *x* sine hyperoymer/hyponymer beregnes som følger: Hvis denotasjonen til *x*s et eller flere opphavstrekk overstiger verdien satt for *SynsetLimit*, oppføres denotasjonene som hyponymer (underbegreper) til *x*. Hyperonymer (overbegreper) beregnes på grunnlag av denotasjonene til *xs* arvede trekk. Hvis denotasjonene til et arvet trekk er større enn *SynsetLimit*, registreres opphavet til dette trekket som hyperonym til *x*.

Synonymer defineres etter følgende tre uavhengige betingelser: (I) Hvis antallet ordbetydninger som arver *x*s såkalte opphavstrekk er mindre enn eller lik *SynsetLimit*, registreres disse denotasjonene som synonymer til *x*. (II) Hvis en ordbetydning *x* har samme kombinasjon av semantiske trekk som en annen ordbetydning *y*, registreres disse som synonymer til hverandre. (I praksis snakker vi da om såkalt *strict synonymy*.) (III) Dersom denotasjonen til et av *x*s arvede trekk er mindre enn eller lik verdien angitt i S*ynsetLimit*, så registreres ordet som ga opphav til dette trekket som synonym til *x*. De øvrige denotasjoner i sistnevnte tilfelle registreres som relaterte ord til *x*, og antas å stå noe fjernere fra *x* enn synonymer. I tesaurusoppslagene presentert i denne oppgaven anvendes en såkalt variabel S*ynsetLimit*, hvor parameterverdien tar hensyn til det totale antall ordbetydninger i et semantisk felt. (Parameterverdien som har blitt funnet å fungere best er å dividere det semantiske feltets antall ordbetydninger på fire).

I neste delkapittel følger en evaluering av betydningsskillene speilmetoden genererer for henholdsvis *rett*N og *rett*A.

2.4 Evaluering av speilmetoden

Dette kapittelet presenterer og evaluerer resultatene for speilmetoden anvendt på henholdsvis *rett*N og *rett*A, basert på tesaurus-oppslagene som speilmetoden genererte. Som sammenligningsgrunnlag ble det valgt å basere seg på den nettbaserte Bokmålsordboka, utviklet ved Universitetet i Oslo[14].

2.4.1 Evaluering av speilmetodens betydningsskiller for "rettN"

Speilmetodens genererte tesaurus-oppslag for *rett*N ble som følger, gitt "default" parameterverdi for *OverlapThreshold* og med variabel *SynsetLimit*.

(18) OverlapThreshold: 0.05
SynsetLimit: variable

rettN
Sense 1
(Translation: course.)
Sense 2
(Translation: court, justification.)
Hyperonyms: argument, plass, ting.
Sense 3
Hyperonyms: krav.
Subsense (i) (Translation: option.)
Synonyms: tilbud.
Subsense (ii) (Translation: rightN.)
Synonyms: adgang, rettighet.
Sense 4
(Translation: dish, food, supper.)
Hyperonyms: kar, fat, måltid, middag, mat.

Den leksikalske enheten *rett*N er som vi ser skilt i fire betydningspartisjoner relativt til engelsk: To betydninger tilknyttet en "matrett"-betydning (betydning 1 og 4) og to "rettslig" relaterte betydninger (betydning 2 og 3). Til sammenligning inndeles *rett*N ifølge Bokmålsordboka ((19) under) i to hovedbetydninger: En "matrett"-betydning og en "rettslig" hovedbetydning som videre sorteres i seks underbetydninger:

[14] Den nettbaserte Bokmålsordboka og Nynorskordboka er tilgjengelig på adressen: http://www.dokpro.uio.no/ordboksoek.html

(19)

TILSLAGSORD	ARTIKKEL FRA BOKMÅLSORDBOKA	
rett	**I rett** m1 (eg sm o s *II rett*, bet. trol fra lty) (kokt el. stekt) mat tillagd for seg el. som del av større måltid *tre r-ers middag / dagens r-* .	
	II rett m1 (norr *réttr*)	
	1	rettighet, lovlig krav *bruksr- / fisker- / fordre, kreve sin r- / stå på r-en sin / ha r- til pensjon / alle har r- til arbeid / være i sin gode r- / komme til sin r-* få bruke sine evner fullt ut, bli verdsatt etter fortjeneste / vederlag, fyllest *gjøre r- og skjell for seg / gjøre r- for maten* gjøre seg fortjent til
	2	rettferdighet, det rette *kjempe for r- og rettferdighet / r- skal være r-* det må gjøres, sies i rettferdighetens navn / *la nåde gå for r-* dømme mildere enn loven egentlig krever / rimelighet, grunn *med en viss r-*
	3	det at noe viser seg å være i samsvar med virkeligheten *ha, få r- i noe / gi en r- i noe*
	4	lov, regel, rettslig vedtak *gjeldende r- / lov og r- / gå r-ens vei* gå til domstolene
	5	domstol *møte for r-en / r-en avsa kjennelse*
	6	i faste uttr i formen *rette*: *stå til r-e* stå til ansvar / *vise, snakke en til r-e* irettesette, tale en til fornuft / *ta seg til r-e* selv ta det en mener en har krav på / *hjelpe en til r-e* veilede en / *finne seg til r-e* tilpasse seg, innrette seg / *sette seg til r-e* sette seg bekvemt og makelig / *forholdene lå vel til r-e* passet, var gode / *komme til r-e* bli funnet / *gå i r-e med en* bebreide en / *med r-e* med god grunn .

Intuitivt kan vi være enig med Bokmålsordboka i at speilmetodens to "matrett"-betydninger 1 og 4 burde blitt slått sammen som én betydning. Grunnen til at de likevel opptrer som to atskilte betydninger ifølge speilmetodens betydningspartisjonering er sparsomme data i parallellkorpuset: Betydningene 1 og 4 kunne bare blitt slått sammen som én betydningspartisjon hvis der fantes et norsk ord i *rett*Ns inverse oversettelsesbilde som korresponderte med både *course* og minst ett annet ord fra *rett*Ns første oversettelsesbilde. (F.eks. hvis *måltid* korresponderte med både *course* i betydning 1 og *dish* i 4, hvilket intuitivt ikke synes utenkelig.) Imidlertid var der altså tilfeldigvis ikke noe ord *x* i norsk (annet enn selve *rett*N) som korresponderte med både (i vårt tilfelle) *course* og et av de engelske ordene som representerer betydning 4.

Dette illustrerer med andre ord en svakhet ved ressursene speilmetoden er basert på, nemlig at jo fattigere data fra parallellkorpuset (i.e. mangel på overlapping mellom oversettelseskorrespondanser), jo flere betydningsskiller blir speilmetoden tvunget til å generere. Selv om speilmetoden altså ikke hadde nok data til å slå sammen betydning 1 med betydning 4, er det likevel tilfredsstillende at speilmetoden lyktes i å samle de øvrige engelske ordene i én betydningspartisjon (betydning 4) som representerer "matrett"-betydningen av *rett*N.).

Vi skal nå se nærmere på de "rettslig relaterte" betydningene 2 og 3, repetert under som (20).

(20) **rettN**
Sense 2
(Translation: court, justification.)
Hyperonyms: argument, plass, ting.
Sense 3
Hyperonyms: krav.
Subsense (i) (Translation: option.)
Synonyms: tilbud.
Subsense (ii) (Translation: rightN.)
Synonyms: adgang, rettighet.

Det er vanskelig å evaluere betydning 2 og betydning 3 opp mot hva man ideelt kunne ønske som betydningsskiller, fordi den "rettslige" betydningen av *rett*N innbefatter flere konsepter hvis grenser er noe flytende. Dette ser vi tydelig med Bokmålsordbokas oppslag nummer II som referansepunkt, hvor det foreslås en underinndeling i seks underbetydninger. Nærmere bestemt kan fire av deres foreslåtte underbetydninger karakteriseres som abstrakte begreper: "et lovlig krav" (oppslag 1); "lov og rett" (oppslag 4); samt to ikke fullt så juridiske begreper som går på "rettferdighet/det rette" (oppslag 2) og om det å "å gi, få eller ha rett i noe i samsvar med virkeligheten" (oppslag 3). De regner dessuten med en konkret "domstols"-betydning (oppslag 5), mens det sjette oppslaget innbefatter faste uttrykk i formen *rette*, som i uttrykket *med rette*.

Speilmetoden genererer på sin side to atskilte hovedbetydninger, gjengitt i (20) som betydningene 2 og 3 av *rett*N, og betydning 3 inndeles videre i to underbetydninger på basis av "default" verdi for parameteren *OverlapThreshold*. Betydning 2 er representert ved oversettelseskorrespondansene *court og justification*, mens betydning 3 tilsvarer betydningspartisjonen som inneholder *claim*, *entitlement, law, option, order* og *rightN*. Intuitivt er det vanskelig å karakterisere betydning 2 som klart distinkt fra nummer 3, verken på grunnlag av de engelske denotasjonene eller ut fra betydningenes respektive synonymer/nærrelasjoner i tesaurusoppslaget.
Betydning 2 heller i retning av den konkrete "domstolsbetydningen" i henhold til Bokmålsordbokas oppslag 5, hvis vi legger oversettelsen *court* og hyperonymet (overbegrepet) *ting* til grunn. I tillegg omfatter speilmetodens betydning 2 et konsept om "å være berettiget" (med utgangspunkt i oversettelseskorrespondansen *justification* og hyperonymet *argument*), i grenseland mellom Bokmålsordbokas begreper nummer 1 og 3.

Som et apropos observerer vi også, litt pussigere, at ordet *plass* er oppført som hyperonym til *rett*Ns betydning 2 - Altså befinner dette ordet seg i samme semantiske felt som en *rett*N_2. Dette viste seg å komme av *rett*Ns oversettelsespartner *court*. Lemmaet *court* er ifølge speilmetoden oppført med fem betydningspartisjoner[15], hvor den største av disse partisjonene inneholder følgende medlemmer: *argument, gård, gårdsplass, plass, rettN, sak* og *ting*. Åpenbart burde disse medlemmene blitt skilt i minst to grupper, nemlig med *argument, rettN, sak* og *ting* som én betydningspartisjon og med *gård, gårdsplass* og *plass* i en annen. Grunnen til at dette ikke slo til er at de to foreslåtte gruppene oversettelsesmessig lenkes sammen av engelsk *point*, som korresponderer med både *plass, sak* og *ting*. Siden *plass* i sin tur videre lenkes til "gårdsplass"-betydningen av *court*, mens ord som *sak* og *ting* likeledes innordnes den "rettslige" betydningen av *court*, gjør oversettelsesbildet til *point* at de foreslåtte betydningsskillet ovenfor i stedet slås sammen til en stor betydningspartisjon. Dette

[15] Innenfor rammene av det manuelt ekserperte materialet i denne hovedoppgaven blir betydningsskiller generert for alle ord inneholdt i det opprinnelige ordets første og inverse oversettelsesbilde; selv om disse betydningsskillene da ikke har like mye informasjon om oversettelseskorrespondanser å gå ut fra som det opprinnelige ordet. Tilgang på ordparallellstilling av ENPC, slik at speilmetoden kan anvendes direkte på parallellkorpuset, vil løse opp denne praktiske begrensningen i tilgang på oversettelsesmateriale.

tilfellet vil bli nærmere diskutert, sammen med et annet tilfelle av uforutsett overlapping mellom korrespondanser, etter presentasjonen av betydning 3 som nå følger.

Speilmetodens betydning nummer 3 synes konseptuelt noe mer enhetlig enn betydning 2. Betydningen representeres av betydningspartisjonen *claim, entitlement, law, option, order* og *right*N i sin totale betydningspartisjon, og har hyperonymene *tilbud*, *adgang* og *rettighet* ifølge tesaurusoppslaget. Dette ligger nært opp til en ganske plausibel kombinasjon av Bokmålsordbokas underbetydninger 1, 2 og 4; som et begrep om "lovlig/rettslig krav og rettferdighet". Karakteristisk for ordene i betydning 3 synes å være at de fanger inn begreper som er relatert til krav/rettferdighet, men ikke nødvendigvis i lovens forstand.

Meningen med speilmetoden er å la oversettelseskorrespondanser i et korpus diktere hvilke betydningsskiller som foreligger. Gitt den flytende grensen mellom rettslige begreper er det vanskelig å argumentere for at sorteringen av konsepter i betydningene nummer 2 og 3 er direkte urimelig. Intuitivt kan man tenke seg at hvis den "rettslige" betydningen av *rett*N skulle deles i to hovedbegreper, så burde skillet gå på et konkret (håndfast) vs. abstrakt rettsbegrep, eller eventuelt at det skilles mellom "lovlig/juridisk" rett vs. begreper som ikke nødvendigvis innebærer rett/rettferdighet i lovens forstand. Det kan således argumenteres for at skillet mellom betydning 2 og 3 faktisk kunne sies å være fullstendig plausibelt om engelsk *justification* hadde blitt tilordnet betydning 3 framfor betydning 2. Da ville man kunne si at betydning 2 er å karakterisere som et konkret og juridisk rettssals-/rettsbegrep, mens betydning 3 fanger inn begreper som ikke *nødvendigvis* utgjør begreper om lov og rett i juridisk forstand: I faktisk språkbruk er det f.eks. mulig å "kreve retten til noe" selv om kravet ikke nødvendigvis er et *de facto* rettslig krav. Likeledes kan også et begrep om "lover" like gjerne referere til naturlover, eller uskrevne normer og regler i samfunnet.

Ved nærmere ettersyn viste det seg da også at *court* og *justification* kom med i samme betydningspartisjon fordi resultatene presentert her er basert på en heuristisk utvidelse av det faktiske ekserperingsmaterialet som forelå i følge parallellkorpus. Bakgrunnen for en heuristisk utvidelse av et ord *x*s mulige oversettelsespartnere er å se på korrespondentene til andre ord i *x*s inverse oversettelsesbilde (i.e. andre ord innenfor samme språk som *x*). Tanken er at hvis tilstrekkelig mange ord i det inverse oversettelsesbildet av *x* kan oversettes til et gitt ord i det motstående språket, så er sannsynligheten stor for at også *x* kan korrespondere med dette ordet, selv om det ikke faktisk foreligger eksempler på en slik korrespondanse i et tross alt begrenset parallellkorpus (for nærmere detaljer, se Dyvik (2002)).

For at to leksikalske enheter (for eksempel *court* og *justification*) skal innordnes en felles betydningspartisjon, må der finnes minst ett ord *x* i det inverse oversettelsesbildet av *rett*N som korresponderte med begge ordene ovenfor. Ifølge det faktiske ekserperingsmaterialet forelå ikke noe slikt ord i korpus (slik at *court* og *justification* følgelig burde separeres i ulike betydningspartisjoner; se (21) nedenfor). Ved en heuristisk utvidelse av ekserperingsmaterialet har imidlertid norsk *argument*, som i følge korpus korresponderer med *justification*, blitt "gjettet" å også kunne korrespondere med *court* selv om en slik korrespondanse ikke forelå i korpus. På dette punktet må nok parallellkorpuset sies å være i samsvar med intuisjonene våre: Det er vanskelig å tenke seg en kontekst hvor engelsk *court* og norsk *argument* kan forekomme som oversettelsespartnere.

Vi ser altså at i tilfellet *court*/*argument* slo kanskje ikke den heuristiske utvidelsen av ekserperingsmateriale helt heldig ut. Det skal likevel nevnes at utvidelsene generelt later til å ha vært vellykket: Basert kun på de oversettelseskorrespondanser som faktisk forelå i korpus, inndeles *rett*N i så mye som åtte betydningspartisjoner ((21) under), hvor det er åpenbart at betydningsskillene er unødvendig atskilt. For eksempel genereres da som vi ser fire ulike betydninger av "matrett"-betydningen, hvor sorteringen konseptuelt sett dessuten ikke synes spesielt velmotivert.

(21) **Sense Partitions of first t-image of "rettN" (Non-extended bases)**
((course)
(court)
(dish)
(justification)
(option)
(claim entitlement law order rightN)
(specialN)
(food supper))

Tilbake til resultatene for de to "rettslige" betydningene basert på utvidet ekserperingsmateriale i (20), kan skillet mellom to "rettslige" betydninger av *rett*N oppsummeres som følger. Siden speilmetodens betydningsskiller synes å antyde at betydning 2 representerer et konkret/juridisk rettsbegrep, kunne det kanskje vært ønskelig at *justification* ble innordnet betydning 3 framfor betydning 2. Dette ville stemme overens med at betydning 3 på sin side fanger inn en mer abstrakt betydning av *rett*N, som innbefatter ord vi kan bruke for å snakke om både krav/rettferdighet/orden i sin alminnelighet, så vel som i juridisk forstand. Gitt de uklare grensene mellom ulike rettslige begreper må det like fullt sies at om man først skal foreslå en korrigering av resultatene med hensyn til kontrastiv flertydighet, ville antagelig det eneste forslaget som tross alt kan forsvares være å slå sammen betydning 2 og 3. Det er nok heller ikke utenkelig at betydning 2 og 3 kunne blitt slått sammen, gitt en større mengde data om oversettelsesmessige korrespondanser enn det som foreligger i ENPC.

Den mest interessante oversettelseskorrespondansen assosiert med betydning 3 er engelsk *order*: Det viste seg at *rett*N kun korresponderte med *order* i én setning i ENPC, nemlig i en setning hvor *rett*N har en betydning relatert til matretter:

(22) *Når en **rett** var klar til servering, ringte madame med en liten klokke på kjøkkenet, og (...)*
(PM1T)
When an order was ready, Madame would clang a bell in the kitchen and (...)

Følgelig skulle faktisk *order* strengt tatt vært inkludert i enten betydning 1 eller 4 (matrettbetydningene av *rett*N). En titt på oversettelseskorrespondentene til *order*, gjengitt i (23) under, viser imidlertid tydelig hvorfor ordet likevel ble innordnet den rettslige betydningen av *rett*N: Substantivet *order* er selv flertydig, og det flertydige ordet *rett* er faktisk det eneste leksemet i oversettelsesbildet som vi "vet" knytter engelsk *order* til en matbetydning. Ordet *lov* i (23) knytter imidlertid *order* til *law*, som jo er en oversettelseskorrespondent til *rett*N. Riktignok kan ordene *bestilling* og *ordre* tenkes å kunne relateres til "bestillingen av en matrett" (på samme måte som *order* selv kom med), men i korpus forelå ingen slike korrespondanser.

(23) **First t-image of "order"**
(orden ordre måte bestilling rekkefølge medalje stand klasse sjikt kommando signal befaling påbud pålegg lov vedtak bestemmelse kjennelse forelegg regel rett system)

Order kan således beskrives som et særtilfelle som setter på prøve antagelsen om at vi ikke forventer at et ord *x* har et sett av betydningsskiller som er identisk med betydningsskillene til et ord *y* i det motstående språket. Flertydigheten til *order* er kanskje ikke er direkte identisk med *rett*N, men deres respektive flertydighet ligger likevel så nært opp til hverandre at det nærmest må betegnes som tross alt "heldig" at *order* ikke hadde en korrespondent som definitivt lenket ordet til en "matrett"-betydning. Hadde *order* f.eks. korrespondert med *middag*, som i: *when the order was served – da middagen var servert*, så

ville *order* gjort at "matrett"-betydningen og den "rettslige" betydningen av *rett*N ble slått sammen i en stor betydningspartisjon.

Som en konklusjon på speilmetodens genererte betydningspartisjoner substantivet *rett*, synes resultatene totalt sett rimelige. Det er synd at betydning 1 og 4 ikke slås sammen til en stor betydning av "matrett", og grunnen til at de skilles som to betydninger bunner i oversettelsesmessige "hull" i korpus. Det har videre blitt argumentert for at skillet mellom de to rettslige betydningene 2 og 3 kan tolkes som å antyde et plausibelt skille mellom et konkret/juridisk vs. et mer abstrakt/allment begrep om rett og rettferdighet. Enkelte av ordene som sorterer under hver av disse betydningene gjør det imidlertid vanskelig å fastslå at det foreligger en klar kontrastiv betydning mellom de to betydningene. Det har blitt påpekt at med mer informasjon om oversettelsesmessig overlapping enn det som forelå i ENPC, er det ikke usannsynlig at betydningene 2 og 3 kunne blitt slått sammen.

Som punkt nummer to har vi også sett to eksempler på at uforutsett overlapping mellom oversettelseskorrespondanser kan skape problemer for speilmetoden. Det første tilfellet forekom i forbindelse med betydning 2 av *rett*N, hvor ordet *plass* noe feilaktig inngår i samme semantiske felt som *rett*N_2. Dette tilfellet illustrerer muligheten for at et flertydig lemma *x* (*court*) har et lemma *y* (*point*) i sitt inverse oversettelsesbilde, hvor *y*s oversettelseskorrespondanser skaper uønskede lenker mellom lemmaer i *x*s første oversettelsesbilde. (mellom lemmaer som selv er flertydige, og som innbyrdes er semantisk urelatert i forhold til *x*). En av antagelsene som ligger til grunn for det å benytte oversettelseskorrespondanser i et parallellkorpus som grunnlag for å utlede enspråklige betydningsskiller for et lemma, er at to lemmaer *x* og *y* ikke forventes å ha den samme type flertydighet (som da ville reflekteres i deres respektive oversettelsesbilder). Et tilfelle som forklart ovenfor kan ikke sies å sette denne antagelsen på prøve. Derimot ser vi at to lemmaer *x* og *y* tilfeldigvis likevel kan ha felles oversettelsespartnere, men på ulikt semantisk grunnlag.

Det andre tilfellet av uforutsett overlapping forekom i *rett*Ns betydning 3. Engelsk *order* ble i ENPC registrert som en oversettelsespartner til *rett*N på grunnlag av en bestillingsbetydning av *order* (som illustrert i setning (22) ovenfor, hvor det dreide seg om en matbestilling). Fordi *order* også kan relateres til et begrep om "lov", ble dette ordet innordnet "gal" betydningspartisjon for *rett*N, selv om det intuitivt nærmest synes mer plausibelt å inkludere *order* under betydning 3 snarere enn under en matrettbetydning. Dette tilfellet illustrerer tydelig uforutsigbarheten i faktisk språkbruk: I utgangspunktet ville man ikke si at *rett*N deler flertydigheten "matrett" og "lov/rett" med *order*, siden man intuitivt heller ville si at det er *bestillingen* av en matrett (snarere enn selve *matretten*) som skaper korrespondansen. Imidlertid ser vi også at korrespondansen mellom ordene i "matrett"-betydningen blir plausibel i den gitte konteksten i setningen (22) ovenfor.

Det første av de to tilfellene forklart ovenfor får ingen direkte konsekvenser for selve betydningsskillene til *rett*N, som er speilmetodens relevante output med tanke på denne oppgavens metode for automatisk betydningstagging. Tilfellet hvor engelsk *order* tilordnes "gal" betydningspartisjon vil imidlertid utgjøre en feilkilde når *rett*Ns betydningspartisjoner brukes som grunnlag for å tagge forekomster av *rett*N i korpuset. Siden der bare finnes et eksempel på at *rett*N og *order* korresponderer i ENPC, burde imidlertid ikke feilkilden gjøre stor skade i praksis.

Hvor store de to uforutsette problemene beskrevet her alt i alt er for speilmetoden, er et empirisk spørsmål som krever en mer omfattende utarbeiding av ekserperingsmateriale til ordnettet for å si noe mer om.

2.4.2. Evaluering av speilmetodens betydningsskiller for "rettA"

Adjektivet *rett* var langt mer frekvent i korpus enn substantivet *rett*, og som forventet kan dette sies å være reflektert i speilmetodens output. Med et mer omfattende ekserperingsmateriale kan man forvente en større overlapping mellom oversettelsesbilder, og følgelig færre (og større) betydningsgrupper (Skjønt man forventer like fullt at kontrastiv flertydighet skilles som ulike betydninger i den grad en slik flertydighet foreligger). Ifølge ekserperingsmaterialet som er input til speilmetoden, deles adjektivet *rett* inn i to hovedbetydninger, som vist i (24) under:

(24) (Extended bases)
OverlapThreshold: 0.05
SynsetLimit: variable

rettA
Sense 1
(Translation: level.)
Sense 2
(Translation: deep, full, immediate, just, rather, rightA, soon, straight.)
Hyperonyms: mye, svær, hel, stor.
Synonyms: ben, direkte, flat, glatt, omgående, rak, rank, snar, snarlig, stiv, straks, tett, tvers, øyeblikkelig.
Related words:
akkurat, felles, fort, full, før, grei, heller, ikke_lenge, korrekt, kort, kun, lik, med_det_samme, med_ett, nettopp, nylig, nøyaktig, om_litt, passe, rask, rett_og_slett, riktig, sen, solid, tidlig, umiddelbar, utelukkende.

Til sammenligning foreslår Bokmålsordboka å sortere adjektivet *rett* i fire underbetydninger; "bein/rank" (oppslag 1), "riktig/korrekt" (oppslag 2) og "riktig/rettferdig" (oppslag 3), mens oppslag 4 inneholder adjektivets adverbiale funksjon. (Hvilket for øvrig passer bra i evalueringen av speilmetodens resultater, gitt avgjørelsen om å slå sammen adjektivet og adverbet *rett* i ekserperingen.) Bokmålsordbokas oppslag er gjengitt i (25) under.

(25)

TILSLAGSORD	ARTIKKEL FRA BOKMÅLSORDBOKA	
rett	**III rett** a2 (norr *réttr*)	
	1	bein, rank *en r- linje / r- til værs / r- i ryggen*
	2	riktig, korrekt *skrive r- / løse oppgaven r- / huske r- / finne det r-e ordet / være på r- vei / i r-(e) tid / se hva som er vrangt og r- / strikke r- og vrangt,* se **vrang*
	3	riktig, rettferdig *finne noe r- og riktig / kjempe for det r-e*
	4	adv: direkte *r- opp, ned / gå r- hjem / r- og slett* simpelthen / like *r- utenfor* / just, akkurat *r- nå / r- som det var* plutselig / svært, riktig *r- lenge / det var r- varmt / r- ofte* .

For først å kommentere sammenligningsgrunnlaget, så er det vanskelig å karakterisere Bokmålsordbokas inndelinger med hensyn til hva motiverer skillene mellom dem. Beskrivelsene er rimelige nok ("bein, rank", "riktig, korrekt" osv), men ser man på en eksemplifisering, som oppslag 1 sitt eksempel *en rett linje*, så er det tross alt mulig å tenke seg at dette eksempelet i en gitt kontekst kunne illustrere "korrekt"-betydningen i oppslag 2. Dette illustrerer at det for *rett*A (som for de fleste adjektiver) foreligger vaghet mellom de ulike konseptene (Dyvik 1998). Slike ords betydningsomfang kan som regel relateres til det som denoteres, slik at spørsmålet mer er hvordan én større, "vag" betydning sorterer mellom mulige kontekster med hensyn til tolkning.

I så måte er kanskje hovedinnvendingen mot speilmetodens to betydningsskiller for adjektivet *rett* at der ikke bare er én stor betydning. Betydning 1, kun representert av det engelske adjektivet *level*, viste seg da også å være utskilt som en egen betydning på grunn av manglende informasjon i korpus: Engelsk *level*A ble kun registrert med én oversettelseskorrespondent i hele korpuset, nemlig *rett*, og følgelig var der ingen mulighet for at ordet kunne overlappe oversettelsesmessig med andre ord. Slike "hull" i korpus er uheldig, men intuitivt kan det slås fast at det ikke er utenkelig at *level* kunne blitt lenket til betydning 2 via f.eks. en oversettelsespartner som norsk *ben*.

Med tanke på å benytte betydningsskillene for en automatisk betydningstagging av korpus innebærer det minimale grunnlaget for betydning 1 at der bare vil finnes én forekomst av *rett* i hele korpuset som korresponderer med det engelske ordet som representerer betydning 1 – Alle øvrige adjektiviske/adverbiale forekomster av *rett* i korpus vil tagges som betydning 2. Vi slår derfor fast at betydningsskillene utledet for adjektivet *rett* ikke er spesielt interessant å benytte for automatisk betydningstagging: Det ville i så fall være mer naturlig å anta en stor betydning av *rett*A, som dermed "disambigueres" på grunnlag av part-of-speech (pos)-tagger. Vi vil derfor heller ikke gå veldig i detalj om betydningsskillene (eller i praksis, den mer informative betydning 2) i det følgende.

Betydning 2 utgjør systematiseringen av et stort semantisk felt som ved (den relativt lave) "default" parameterverdien for *OverlapThreshold* ikke deles inn i underbetydninger. Med variabel parameterverdi for *SynsetLimit* registreres fire overbegreper for denne betydningen, nemlig *mye, svær, hel* og *stor*. Gitt ordbetydningens inndeling mellom synonymer og relaterte ord når der ikke er noen underbetydninger, synes resultatet fornuftig: Synonymene, som antas å stå nærmere *rett*A_2 semantisk enn de relaterte ordene, ville enkelt kunne substitueres med *rett* i en gitt kontekst. Eksempler: *En rett/ben/flat vei; en rett/rak/rank/stiv rygg; Jens gikk rett/direkte/omgående hjem,* etc. De fleste relaterte ordene er derimot ikke like enkle å substituere, likevel kan vi være enige at de semantisk befinner seg i samme "nabolag" som *rett*A. Disse ordbetydningene går stort sett på "korrekthet/nøyaktighet"

(kanskje tilsvarende Bokmålsordbokas oppslag 2), "nylighet" og "kortvarighet" (som kan lenkes til tidsadverbialet *rett*).

Om vi anvender "default" *SynsetLimit* (Dyvik, 2002) og øker *OverlapThreshold*-parameteren fra "default" verdi 0.05 til 0.75 slik at flere underbetydninger genereres, ser vi av (26) under at fire underbetydninger genereres for betydning 2:

(26) Extended bases
OverlapThreshold: 0.75
SynsetLimit: 20

rettA
Sense 1
(Translation: level.)
Sense 2
Hyperonyms: hel, mye, stor, svær.
Subsense (i) (Translation: straight.)
Synonyms:
ben, direkte, flat, glatt, omgående, rak, rank, snarlig, stiv, tvers, øyeblikkelig.
Subsense (ii) (Translation: soon.)
Synonyms: snar
Related words:
fort, før, glatt, heller, ikke_lenge, kort, lik, med_det_samme, med_ett, om_litt, omgående, rask, sen, snarlig, straks, tidlig, øyeblikkelig.
Subsense (iii) (Translation: rightA)
Synonyms: tett.
Related words:
akkurat, grei, korrekt, kun, lik, nettopp, nylig, nøyaktig, passe, rett_og_slett, riktig, solid, tvers, utelukkende.
Subsense (iv) (Translation: immediate)
Synonyms: straks.
Related words:
ben, direkte, felles, full, glatt, ikke_lenge, kort, lik, med_det_samme, med_ett, omgående, snarlig, umiddelbar, øyeblikkelig.

Vi ser at alle underbetydninger inneholder synonymer/relaterte ord som uttrykker et tidsaspekt, f.eks. *øyeblikkelig, omgående* og *snarlig*. Underbetydningene (ii) og (iv) kan ganske klart karakteriseres som underbetydninger relatert til en betydningsnyanse av *rett* som går på "nylighet/kortvarighet". De to øvrige underbetydningene er litt vanskeligere å tilskrive en enhetlig betydningsnyanse. Med unntak av synonymer/ord relatert til tid i underbetydningene (i) og (iii), gjør synonymer som *direkte, rak*, *rank* og *stiv* det nærliggende å jevnføre (i) med Bokmålsordbokas betydningsnyanse 1; "bein/rank". I tillegg kan vi kanskje legge til "flat/glatt" som nøkkelord i karakteristikken, mens (iii) mer går på "korrekthet".

En overlappingsterskel på så mye som 0.75 gjør kanskje at flere underbetydninger genereres enn vi ønsker, men samtidig er det fristende å si seg enig i tendensen til stor overlapping mellom de resulterende nyansene som illustreres i (26): Bokmålsordbokas inndeling i underbetydninger for adjektivet *rett* er forsåvidt klart nok definert, men vi aner likevel at de ulike betydningsnyansene ikke er like klart atskilt i faktisk språkbruk. Det er som sådan slående at heller ikke Bokmålsordboka har definert noen egen betydningsnyanse for tidsaspektet som *rett* kan ha. I stedet inngår tidsaspektet ganske enkelt sammen med *rett*A sine adverbiale funksjoner: Bokmålsordboka gir da eksempler på "nylighets"-aspektet ved synonymene "just, akkurat" og "plutselig".

Som en konklusjon kan vi si at adjektivet *rett* vanskelig kan sies å inneha kontrastiv flertydighet, snarere representeres ulike betydningsnyanser med en ganske høy grad av overlapping. For denne oppgavens formål med å lage en betydningstagger er derfor ikke resultatene for *rett*A videre aktuelle å bruke.

2.5 Oppsummering og konklusjon

Dette kapittelet har tatt for seg første steg i oppgaven med å lage automatisk betydningstagger, nemlig å benytte speilmetoden for hente ut de betydningsskiller som foreligger for et gitt lemma i henhold til oversettelseskorrespondanser i parallellkorpuset ENPC.

Det første som da måtte gjøres var å registrere oversettelseskorrespondanser i ENPC manuelt. For å treffe mest mulig konsekvente valg under ekserperingsarbeidet var det nødvendig å utarbeide prinsipper som kan avgjøre hva som faller inn under begrepet "oversettelsesmessig korrespondanse". To hovedhensyn motiverte formuleringen av prinsippene: For det første er det ikke ønskelig å registrere tilfeller hvor det intuitivt synes klart at der ikke er en god korrespondent til søkeuttrykket i den parallellstilte setningen, da dette kunne utgjøre en unødvendig feilkilde for sorteringen av ordbetydninger. På den andre siden ønsker vi heller ikke å være for strenge, siden vi ønsker å hente ut mest mulig informasjon fra korpuset. Hvorvidt vi har lyktes i sistnevnte hensyn kan kanskje bedre besvares ved å se på resultatene fra den automatiske betydningstaggeren: I hvilken grad har betydningstaggeren tilstrekkelig informasjon for å tagge flest mulig forekomster av et lemma hvor vi som mennesker ganske umiddelbart kunne fastslått betydningen?

Med hensyn til hvorvidt ekserperingsprinsippene åpner for feilkilder i oversettelsesmaterialet, det vil si hvorvidt ekserpringsprinsippene kanskje ikke var strenge nok, er det rimelig å si at ekserperingsprinsippene har fungert bra. I den grad betydningsskillene til et lemma ikke er helt som ønsket eller et semantisk felt inneholder uønskede ordbetydninger (f.eks. inklusjonen av *plass* sammen med en "rettslig" betydning av *rett*N), kan dette vanskelig relateres til ekserperingsprinsippene. Problemene later ikke til å bunne i selve oversettelseskorrespondansen, snarere kommer det dels av fattige data i parallellkorpuset og dels av uforutsigbarhet med hensyn til hvordan lemmaer tilfeldigvis kan lenkes sammen oversettelsesmessig.

Betydningsskillene som utledes for *rett*N og *rett*A kan alt i alt sies å være tilfredsstillende. Det har blitt argumentert for at adjektivet *rett* sine to betydningsskiller er å regne som en følge av mangelfull informasjon om oversettelsesmessige korrespondanser i parallellkorpuset. Betydning 1 er denotert av ett engelsk ord, som kun korresponderer med *rett*A én gang i ENPC, og som intuitivt ikke overhodet er kontrastivt flertydig i forhold til den store betydningen 2. I praksis kan man dermed heller snakke om én stor betydningspartisjon, som synes rimelig i forhold til hva vi lingvistisk kunne ønske. Med tanke på den automatiske betydningstaggeren er resultatene for *rett*A imidlertid ikke relevant å anvende, siden bare én forekomst av *rett*A i hele korpuset ville kunne tagges som betydning 1 mens alle øvrige forekomster ville tagges som betydning 2.

Substantivet *rett* sine betydningspartisjoner synes derimot anvendbare som utgangspunkt for å semantisk tagge forekomster av *rett*N i korpus. Det har blitt påpekt at "matrett"-betydningen 1, kun representert ved engelsk *course*, med fordel kunne blitt slått sammen med den større "matrett"-betydningen nummer 4. Det er også mulig at også de to rettslig relaterte betydningene 2 og 3 burde bli slått sammen. Skillet mellom de to sistnevnte synes likevel plausibelt til en viss grad.

Av resultatene speilmetoden gir for *rett*N og *rett*A kan vi slå fast at oversettelsesmessige "hull" som følge av størrelsen på parallellkorpuset er et problem. De klareste eksemplene på dette er substantivet og adjektivet *rett* sine respektive betydninger nummer 1: Sannsynligvis ville begge disse betydningspartisjonene, som bare var representert ved ett engelsk ord hver (substantivet *course* og adjektivet *level*), blitt slått sammen med en annen og større betydningspartisjon om ekserperingsmaterialet hadde vært mer omfattende.

En "mindre" betydningspartisjon som *rett*Ns betydning 1, som vi intuitivt aner burde blitt innordnet en større betydningspartisjon, er noe uheldig med tanke på benytte betydningsskillene som grunnlag for å semantisk tagge forekomstene av *rett*N i parallellkorpuset. Dette semantisk taggede korpuset skal gis som input til et maskinlæringssystem for orddisambiguering, og ideelt bør maskinlæringssystemet ha flest mulig eksempler på bruken av hver ordbetydning. Maskinlæringssystemet vil i tilfellet *rett*N måtte prøve å lære et skille som egentlig ikke eksisterer mellom de to matrett-betydningene, hvilket også innebærer at der blir færre eksempler på bruken av hver betydning. Selv om betydningsskillet mellom betydning 2 og 3 synes noe mer fornuftig, har det blitt observert at overlappingen mellom dem konseptuelt sett er såpass påfallende at de sannsynligvis kunne blitt slått sammen i én stor betydningspartisjon hvis oversettelseskorrespondansene fra parallellkorpuset var mer omfattende.

I neste kapittel presenteres programmet som automatisk tagger forekomster av *rett*N semantisk, basert på de fire betydningspartisjonene som er blitt presentert.

3 Automatisk ekstrahering av et betydningstagget korpus

3.1 Innledning

Forrige delkapittel viste at speilmetoden utleder fire oversettelsesbaserte betydningspartisjoner for substantivet *rett*, gitt lemmaets oversettelseskorrespondanser ifølge parallellkorpuset ENPC. Disse er gjengitt i (1) under. Det ble videre konkludert at disse betydningsskillene gjør *rett*N til et adekvat testlemma for forsøket på å benytte speilmetodens betydningspartisjoner for å semantisk tagge forekomster av lemmaet i korpus.

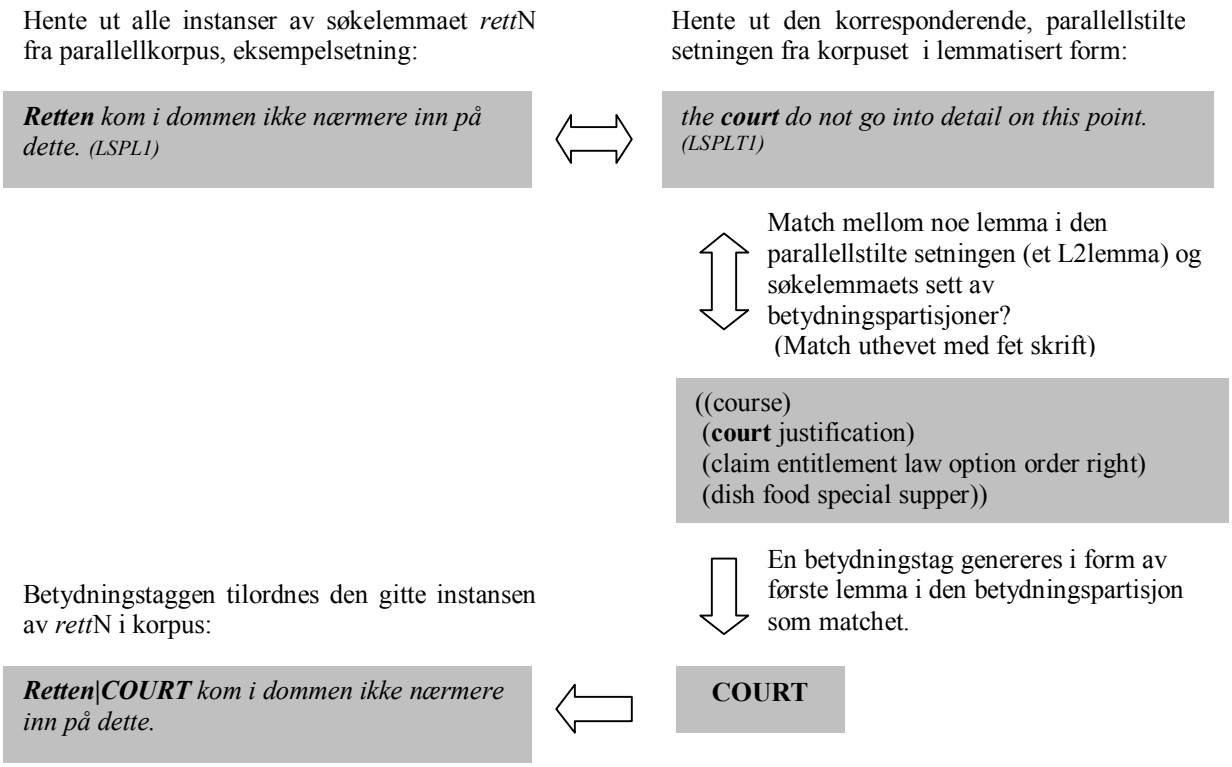
(1) **Sense partitions of first t-image of "rettN":**
((course)
(court justification)
(claim entitlement law option order rightN)
(dish food specialN supper))

Dette delkapittelet presenterer en uovervåket metode for å ekstrahere et finitt, betydningstagget korpus på basis av oversettelseskorrespondanser i et parallellkorpus, slik oversettelseskorrespondansene er sortert betydningsmessig ved speilmetoden. Vi vil referere til metoden som en metode for *Automatisk Betydningstagging av et Treningskorpus*, som også kan forkortes til ABT. Forsøket er motivert av problemet med manglende tilgang på semantisk taggede treningskorpora til bruk innen maskinlæringsmetoder for WSD. Formålet med forsøket er å undersøke i hvilken grad oversettelseskorrespondanser i parallellkorpus, sortert etter betydning ved speilmetoden, er egnet som kilde for å automatisere betydningstaggingen av et treningskorpus som bakenforliggende ressurs for overvåket WSD.

Den presenterte ABT-metoden legger til grunn at hvis speilmetodens oversettelsesbaserte betydningspartisjoner for et ord *x* reflekterer ulike betydninger av *x* i forhold til et annet språk, så kan hver betydningspartisjon tolkes som (et korrelat til) en semantisk kategori av *x*. Ut fra *rett*Ns betydningspartisjoner gjengitt i (1) vil vi altså tolke *rett*N som å ha fire semantiske kategorier. Tanken er å anvende oversettelseskorrespondansene i et lemma *x*s sett av betydningspartisjoner som en indikator på hvilken betydning en gitt instans av *x* har i sin aktuelle kontekst. Metoden er skjematisk illustrert i fig. 3.1 under med testlemmaet *rett*N som eksempel.

ABT-metodens første steg er å søke gjennom hele parallellkorpuset etter instanser av *rett*N. Eksempelinstansen i figuren under er gjengitt i sin setningskontekst for illustrasjon. For hver instans ekstraheres den korresponderende setningens lemmaer (Siden ENPC foreløpig kun er parallellstilt på setningsnivå). Disse vil fra nå av refereres til som *L2lemmaer*. Hvert L2lemma sammenlignes så mot lemmaene i speilmetodens genererte sett av betydningspartisjoner for *rett*N: Hvis et L2lemma matcher et lemma i en betydningspartisjon P_i, genereres en betydningstag i form av det første lemmaet i P_i. Denne tilknyttes så den aktuelle setningsforekomsten av *rett*N. På denne måten konverteres et subsett av parallellkorpuset til et enspråklig, finitt betydningstagget korpus som kan anvendes som treningskorpus for overvåkede maskinlæringsmetoder.

Fig. 3.1
Prosedyren for automatisk betydningstagging, eksemplifisert ved det norske substantivet *rett*



Metoden for automatisk betydningstagging av et treningskorpus ble implementert i LISP (Allegro), og prosesserer et LISP-bilde som representerer ENPC i form av XML-dokumenter. Del (3.2) gir en kort forklaring av de elementene av XML-oppbygningen til parallellkorpuset som er relevante for implementeringen av en automatisk betydningstagger. Del (3.3) tar for seg implementeringen av betydningstaggeren. Selve programkoden er lagt ved som appendiks 1. (3.4) gir en manuelt basert evaluering av ABT-metodens resultater med substantivet *rett* som testlemma. Evalueringen er utført både med hensyn til de betydningstaggede instansene og til de som ble forkastet, i lys av spørsmålet om hvorvidt oversettelseskorrespondanser er anvendbare som grunnlag for automatisk betydningstagging. I (3.5) følger en diskusjon og konklusjon.

3.2 Oppbygningen av parallellkorpuset i XML

Utgangspunktet for betydningstaggerens programkode er et LISP-bilde av parallellkorpuset ENPC i XML-format, laget av Sindre Sørensen ved HIT-senteret. XML-formatet muliggjør å strukturere innholdet hierarkisk, sammenlignbart med en trestruktur, med en dokumentrot (selve korpuset ENPC) og et rotelement/en rotnode som er modernode til alle underordnede nivåer av noder. Innholdet i dokumentet defineres hierarkisk som klasser av objekter, slik at man kan traversere nedover i hierarkiet fra nivå til nivå.

Jeg vil ikke her gå i detalj om den totale trestrukturen, men bare gi et overblikk over de klasser som er relevante for mitt formål. Parallellkorpuset ENPC består av dokumenter; et dokument for hver prosa- eller saklitterære tekst som korpuset har et utdrag fra. I XML-formatet representeres hvert dokument som et *dokumentobjekt* i klassen av dokumenter. Hvert slikt dokumentobjekt har en unik id. Et dokumentobjekt inneholder videre strukturerte *setningsobjekter*: Disse defineres i henhold til to klasser. Objektene i klassen *COMPRESSED-PREALIGNED-SENTENCE* inneholder informasjon om dokumentet sine respektive setninger. Et eksempel på et setningsobjekt er gjengitt i (2) under, hvor setningsobjektet som vi ser er identifisert som *@ #x8162917a*. Hvert slikt setningsobjekt er lenket til sin én eller flere korresponderende setninger i motstående språk. Sistnevnte setningsklasse er kalt CORRESPONDING-SENTENCES, og de er oppført som datternode til setningen den korresponderer med (datterelement nummer 5 i (2)). Den korresponderende setningen utgjør da selv en COMPRESSED-PREALIGNED-SENTENCE i det motstående språk, hvis CORRESPONDING-SENTENCES peker tilbake på @ #x8162917a illustrert under.

(2) Eksempel på et objekt i klassen COMPRESSED-PREALIGNED-SENTENCE[16]

```
COMPRESSED-PREALIGNED-SENTENCE @ #x8162917a
 #<COMPRESSED-PREALIGNED-SENTENCE @ #x8162917a
   0 Class -------- #<STANDARD-CLASS COMPRESSED-PREALIGNED-SENTENCE
   1 TOKEN-ARRAY -- The symbol :--UNBOUND--
   2 SENTENCE-ARRAY - simple T vector (12) = #((() ...)...)
   3 NEXT-SENTENCE - The symbol NIL
   4 PREV-SENTENCE - #<COMPRESSED-PREALIGNED-SENTENCE @ #x816370a2
   5 CORRESPONDING-SENTENCES - (#), a proper list with 1 element
   6 SENTENCE-SCORES - The symbol :--UNBOUND—
```

De sentrale elementene i en slik setningsvektor for implementeringen av en ABT-metode er altså elementet nummer 5, CORRESPONDING-SENTENCES, og i tillegg elementet nummer 2, SENTENCE-ARRAY. Klassen SENTENCE-ARRAY er objekter strukturert i form av en vektor som inneholder setningens ord, representert som *tokens*. Et *token* er en liste med to elementer, med selve ordformen (som forekommer i teksten) som første element og med ordformens tilhørende lemma og pos-tag som andre element. Setningsobjektet i (2) ovenfor sin SENTENCE-ARRAY er gjengitt i (3), hvor vi ser at vektoren består av 12 elementer.

[16] Elementene 3 og 4, NEXT-SENTENCE og PREV-SENTENCE, er del av denne oppgavens implementering. Det ble valgt å innføre eksplisitte lenker mellom den "nåværende" setningen og dens henholdsvis forrige og neste setning for å gjøre programkoden mer oversiktlig. Lenkene til forrige og neste setning anvendes i forbindelse med ekstrahering av kontekst for det flertydige ordet (kap. 4).

(3) SENTENCE-ARRAY for setningsobjektet i (2) ovenfor:

```
A simple T vector (12) @ #x8165d0f2
   0- (() ( . )), a proper list with 2 elements
   1- ("Sagnene" ("sagn" . SUBST)), a proper list with 2 elements
   2- ("har" ("ha" . VERB)), a proper list with 2 elements
   3- (#1="dessuten" (#1# . ADV)), a proper list with 2 elements
   4- (#1="mange" (#1# . ADJ)), a proper list with 2 elements
   5- ("betydninger" ...), a proper list with 2 elements
   6- ("-" ("$-" . <STREK)), a proper list with 2 elements
   7- (#1="de" (#1# . PRON)), a proper list with 2 elements
   8- ("uttrykker" ...), a proper list with 2 elements
   9- ("fortellernes" ...), a proper list with 2 elements
  10- ("holdninger" ...), a proper list with 2 elements
  11- ("." ("$." . CLB)), a proper list with 2 elements
```

Hvis preprosesseringen av en ordform (m.h.t. lemmatisering og pos-tagging) ikke lyktes i å disambiguere mellom flere tolkningsmuligheter, blir alle alternativer listet opp. Som eksempel vises i (4) under formatet hvis pos-taggingen var tvetydig:

(4) ("ordform" ("lemma1" . pos-tag1) ("lemma1" . pos-tag2))

I neste delkapittel følger en redegjørelse av selve programmet for automatisk betydningstagging.

3.3 Implementeringen av en automatisk betydningstagger

Dette kapittelet gir en nærmere presentasjon av programmet som tagger instanser av et gitt lemma *x* semantisk, på grunnlag av speilmetodens genererte betydningspartisjoner. Selve programkoden, skrevet i LISP Allegro (vedlagt i appendiks 1), vil ikke bli diskutert. I stedet presenteres programmet i henhold til de ulike stegene i prosesseringen. Underveis vil det pekes på maskinelle problemer og løsningene vil bli diskutert.

3.3.1 Input

Programmet tar fire argumenter som input:

- Søkelemma
- Søkelemmaets sett av betydningspartisjoner, generert ved speilmetoden
- Ønsket pos-tag for søkelemmaet
- Tilsvarende pos-tag i det motstående språket

Søkelemmaet angir hvilket lemma systemet skal søke etter i korpus for å hente ut instanser for betydningstagging.

Med hensyn til betydningspartisjonene kan det bemerkes at lemmaene i settet av betydningspartisjoner i enkelte tilfeller ble manuelt "pos-tagget" under ekserperingsarbeidet, ved å legge til ordkategori etter lemmaet, e.g. *special*N. Dette ble gjort for å kunne skille

oversettelseskorrespondansene til f.eks. *right*N fra *right*As korrespondanser. "Pos-tagger" som er lagt til lemmaet må tas bort før settet av betydningspartisjoner gis som input til betydningstaggeren, siden lemmaet ellers ikke ville kunne sammenlignes mot L2lemmaer i den parallellstilte setningen.

Som forklart i presentasjonen av parallellkorpuset ENPC (2.2) ble den norske og den engelske siden av parallellkorpuset ENPC syntaktisk tagget ved hjelp av ulike taggingsalgoritmer: Den norske siden er tagget ved hjelp av Oslo-Bergen-taggeren, mens den engelske siden benytter Penn TreeTagger. Siden de to pos-taggingsalgoritmene har ulikt inventar av pos-tagger, må den ønskede pos-taggen spesifiseres i ABT-metodens input for begge språksidene av parallellkorpuset. Søkelemmaets pos-tag må være med fordi adskilte ordkategorier med samme oppslagsform (slik som *rett*) har adskilte betydningspartisjoner, derfor er kun én pos-tag relevant. På samme måte oppføres det motstående språkets tilsvarende pos-tag, siden også lemmaer på den engelske siden kan gå på tvers av ordkategorier. (f.eks. tilfellene *right*N – *right*A og *special*N – *special*A i *rett*Ns sett av betydningspartisjoner). Ved å utnytte pos-informasjon på begge sider av korpus, sikres at systemet kun registrerer en match mellom et lemma og en betydningspartisjon hvis det matchende lemmaet har ønsket pos-tag.

3.3.2 Søkekriterier

Første steg i prosesseringen er å søke seg frem til alle instanser av søkelemmaet i korpus. Dette steget dreier seg altså om å hente ut *kandidater* for betydningstagging – Hvorvidt de tagges eller forkastes avgjøres av selve betydningstaggingskomponenten (3.3.3).

Søket foregår ved at det søkes gjennom alle dokumenter på den aktuelle språksiden av parallellkorpuset. (I vårt tilfelle, på den norske siden.) For hvert dokument traverseres alle dokumentets setninger (representert ved *sentence-array*s), og for hver setning gjennomgår systemet alle *tokens*. Et *token* er som forklart i (3.2) oppbygd etter følgende mønster, hvor stjernen er å tolke som en *kleene star* : ("ordform" ("lemma" . POS-tag)*)

Programmet stiller for hvert *token* følgende to betingelser, som begge må være oppfylt for videre prosessering. Hvis betingelsene ikke slår til, fortsetter systemet til neste *token*. Betingelsene er som følger:

- Inneholder det aktuelle *token* et lemma (av mulige flere oppførte alternativer) som tilsvarer søkelemmaet spesifisert i input?
- Inneholder det aktuelle *token* en pos-tag (av mulige flere oppførte alternativer) som tilsvarer pos-taggen spesifisert i input?

Dette er ganske vide betingelser. Det tillates i praksis at det aktuelle *token* består av flere lemmaer og/eller pos-tagger som vi *ikke* er interessert i, så lenge vårt lemma og pos-tag var oppført som et av alternativene. Grunnen til de vide betingelsene er at preprosesseringen i tilfellet *rett*N slo noe uheldig ut, på tross av at Oslo-Bergen-taggeren ifølge tester skal ha en performans nær 100% (se 2.2). ENPC inneholder 396 instanser av *rett*N, men majoriteten av disse ble ikke disambiguert under lemmatisering/pos-tagging: Ved søk på kun entydig preprosesserte instanser får systemet bare 109 treff[17]. De øvrige forekomstene av *rett*N er

[17] Det viste seg ved manuell gjennomgang at tre av instansene entydig preprosessert som *rett*N egentlig var galt kategoriserte instanser av adjektivet *rett*. Antallet entydige og *riktig* preprosesserte instanser av *rett*N er dermed 106.

tildelt enten flere alternative lemmaer eller flere pos-tagger. For eksempel er så mange som nesten ¼ av instansene oppført med både *rett* og *rette* som mulige lemmaer. Det er vanskelig å si i hvilken grad preprosesseringsproblemene tilfeldigvis var spesielt store for testlemmaet *rett*N, hvor det blant annet antagelig kan spille en rolle at oppslagsformen er felles på tvers av ordkategorier.

På grunn av den uventet usikre preprosesseringen av korpus er det altså ønskelig å tillate at både lemmatiseringen og pos-taggingen er flertydig, for å sikre at flest mulig instanser kommer med. En uheldig konsekvens av de vide søkekriteriene er imidlertid at de åpner for at *uønskede* lemma kan inkluderes blant treffene som systemet finner. For eksempel inkluderes instanser av adjektivet *rett* hvis en slik instans var lemmatisert som *rett* og pos-tagget som ”enten substantiv eller adjektiv”. I ENPC er bøyningsformen *rett* pos-tagget som "enten substantiv eller adjektiv" i 337 tilfeller, hvorav over halvparten er instanser av *rett*A. Alt i alt viste det seg at med *rett*N som testlemma gjør de åpne søkekriteriene at så mange som 1/3 av systemets funnede treff egentlig var forekomster av adjektivet *rett*, samt noen tilfeller av verbet *rette*.

To faktorer gjør likevel at det ble valgt å tillate såpass vide betingelser. For det første illustrerer testlemmaet *rett*N at hvis kun entydig preprosesserte instanser ble vurdert, ville systemet bare funnet ca. ¼ (109 stk.) av de faktiske instansene av søkelemmaet som finnes i korpus. En annen faktor er at treffene som systemet finner i korpus er å anse som *kandidater* for betydningstagging. Hvorvidt en instans faktisk blir semantisk kategorisert, avhenger av at der foreligger en oversettelseskorrespondent som er inneholdt i søkelemmaets sett av betydningspartisjoner. Følgelig kan et uønsket lemma kun inkluderes i det betydningstaggede materialet for *rett*N hvis et tilfeldig ord i den korresponderende setningen gir grunnlag for å generere en substantivisk relatert betydningstag. Veier man denne risikoen mot de få kandidatene systemet ellers ville finne i korpus, syntes det velmotivert å ta denne sjansen. Som vil bli sett, hendte det relativt sjelden at et uønsket lemma feilaktig tagges med en av søkelemmaets betydningstagger - dvs. sjelden relativt til det totale antallet av treff på kandidater som ikke *er* forekomster av søkelemmaet. Som vi vil se i den manuelt baserte evalueringen i (3.4), er det likevel uønsket inkluderte adjektiver som utgjør den store feilkilden i det automatisk betydningstaggede materialet.

Hver slik funnede kandidat for betydningstagging sendes så videre i systemet til komponenten som avgjør om instansen kan tagges for betydning.

3.3.3 Betydningstaggingskomponenten

Denne komponenten kan inndeles i to subfaser. Den første av disse består i å generere potensielle betydningstagger for hver gitte forekomst av søkelemmaet. Siden ENPC foreløpig kun er parallellstilt på setningsnivå, må systemet nødvendigvis søke gjennom hele den korresponderende setningen for å lete etter betydningstagger. Det kan da forekomme at mer enn én betydningstag genereres. (For eksempler, se punkt 2 under). I slike tilfeller går systemet videre til en subfase to som foretar et valg mellom de alternative betydningstaggene, basert på informasjon om setningsposisjon. Under følger en nærmere beskrivelse av de to subfasene.

1. Genereringen av potensielle betydningstagger

For hver kandidat for betydningstagging som systemet finner i korpus, henter systemet ut kandidatens én eller flere korresponderende setninger i det motstående språket, som vi kan referere til som *L2sentences*[18]. For hvert L2lemma i *L2sentences* stiller systemet følgende to betingelser, som begge må være oppfylt for å generere en betydningstag på grunnlag av det gitte L2lemmaet:

- Matcher L2lemmaet et ord i en av betydningspartisjonene $\{P_1, \ldots, P_n\}$ til søkelemmaet *x*?
- Stemmer den syntaktiske taggen til L2lemmaet overens med den eller de pos-taggene som i input er oppgitt å tilsvare søkelemmaets pos-tag?

Hvis begge disse betingelsene er oppfylt, genereres en betydningstag, i form av det første lemmaet i den betydningspartisjonen P_i som L2lemmaet matchet med. Systemet registrerer også L2lemmaets relative setningsposisjon i tilfelle flere enn én betydningstag ble generert. Dette kalkuleres enkelt ved formelen i (5) under.

(5)

$$relativ\ \ setningspo\ sisjon(L2lemma) = \frac{L2lemmas\ \ absolutte\ \ setningspo\ sisjon\ \ n}{Setningsle\ ngde\ \ m}$$

2. Valg mellom flere potensielle betydningstagger

Siden ENPC foreløpig ikke er ferdig parallellstilt på ordnivå, kan det forekomme at det genereres mer enn en betydningstag, slik at systemet må velge mellom dem. (6) illustrerer et tilfelle hvor søkelemmaet opptrer mer enn én gang innenfor samme setning. Det kan også inntreffe at den korresponderende setningen inneholder et eller flere L2lemmaer som *ikke* korresponderer med søkelemmaet, men som tilfeldigvis matcher søkelemmaets betydningspartisjoner. Dette illustreres av eksempelsetning (7), hvor det er betydningstag nummer to av de tre genererte taggene som skal tilordnes instansen av *rett*N. I eksemplene under angis den resulterende betydningstaggen (adskilt med skillelinje og store bokstaver) som genereres for hvert av L2lemmaene som matcher en av søkelemmaets betydningspartisjoner.

(6) *Man har særlig søkt å beskytte foreldrenes **rett** til å ha sine barn hos seg og å være sammen med dem, og dessuten tatt hensyn til foreldrenes **rett** til å bestemme vedrørende sine barns verdslige utdannelse og religiøse oppdragelse.*
(LSPL1)
*Special attempts have been made to protect the parents' **right|CLAIM** to keep their children with them, and account has also been taken of the parents' **right|CLAIM** to decide on the secular education and religious upbringing of their children.*

[18] En setning kan selvsagt korrespondere med mer enn én setning i det motstående språket, siden oversetteren kan ha valgt å splitte opp en setning i to atskilte setninger.

(7) *Da moren fødte tvillinger, nektet mannen å vedta forelegg for tvilling nr 2 — **retten** kom imidlertidig til at det første forelegg måtte omfatte begge barna, selv om entallsformen var brukt.*
(LSPL1)
*When the mother gave birth to twins, he refused to accept the **order|CLAIM** for the second child, but the **court|COURT** found that the first affiliation **order|CLAIM** must apply to both children, even if it had only used the singular form.*

Åpenbart vil man i tilfeller av mer enn én betydningstag være interessert i den betydningstaggen som ble generert på grunnlag av selve søkelemmaets oversettelseskorrespondent. Uten tilgang på eksplisitt ordparallellstilling ble slike tilfeller løst på heuristisk vis, ved å basere seg på relativ setningsposisjon (jf. formelen i (5) ovenfor) for å selektere den antatt riktige betydningstaggen. For hvert $L2lemma_i$ som ga grunnlag for å generere en betydningstag kalkuleres differansen mellom den relative setningsposisjonen til $L2lemma_i$ og instansen av søkelemmaet sin relative posisjon. Jo lavere differanse, jo mer sammenfallende er setningsposisjonene til henholdsvis instansen av søkelemmaet og $L2lemma_i$. Systemet velger den betydningstaggen som ga lavest differanse i forhold til søkelemmaets posisjon.

Det er klart at man ikke har noen garanti for at for eksempel oversetteren har tatt seg oversettelsesmessige friheter med hensyn til setningselementenes plassering. Uten tilgang på ordparallellstilling antar vi likevel generelt at den "nærmeste" betydningstaggen *mest* sannsynlig vil være den riktige. Som vil bli sett i evalueringen, løser informasjon om relativ setningsposisjon opp de aller fleste tilfeller hvor mer enn én betydningstag ble generert.

I neste delkapittel presenteres og evalueres resultatene for betydningstagging av substantivet *rett*, gitt de fire betydningspartisjonene som er generert av speilmetoden.

3.4 Evaluering av den automatiske betydningstaggeren

Wilks (2003) påpeker at den viktigste faktoren for å evaluere et system for orddisambiguering er systemets evne til å produsere korrekte tagger. Andre faktorer, som bruk av maskinelt minne og tid kan være relevante i forbindelse med bruk av systemet innenfor en "større" NLP-applikasjon som maskinoversettelse; altså når maskinell effektivitet i større grad er et spørsmål om brukervennlighet. Siden den automatiske betydningstaggeren presentert i denne oppgaven ikke er ment å inngå som direkte ressurs for et større NLP-system, vil evalueringen derfor fokusere på betydningstaggerens kvalitet med hensyn til korrekte tagger.

Dette delkapittelet setter fokus på hvor godt egnet parallellkorpus og speilmetoden er som ressurs for å utføre selve betydningstaggingen. I (3.4.1) analyseres derfor det betydningstaggede materialet med hensyn til hvilke feilkilder som foreligger i materialet. Deretter følger i (3.4.2) en diskusjon av instansene som *ikke* ble betydningstagget, og det tas stilling til hvordan resultatene belyser anvendeligheten til ressursene som er brukt.

Tabell 3.2 under gir et innledende overblikk over den prosentvise fordelingen mellom betydningstaggede og ikke taggede (forkastede) instanser med *rett*N som testlemma.

Tabell 3.2: Relativ frekvensfordeling mellom antallet instanser som ble betydningstagget og de som ble forkastet. (Tall basert på det manuelt kontrollerte, faktiske antall instanser av *rett*N i korpus angis i parentes)

Betydningstaggede instanser	.35 (.58)
Forkastede instanser	.65
Antall instanser av *rett*N i korpus totalt	646 (396)

Det er verdt å merke seg at korpus basert på manuell kontrollering inneholder 396 instanser av *rett*N, mens systemet finner hele 646 (antatte) treff på søkelemmaet. Denne differansen bunner i de åpne søkekriteriene det ble valgt å benytte: Drøyt 1/3 av systemets *antatte* treff på instanser av søkelemmaet er i virkeligheten instanser av *rett*A og verbet *rette*. For enkelhets skyld vil instanser av andre lemmaer enn søkelemmaet fra nå av refereres til som *uønskede instanser*. Om vi forholder oss til det virkelige antallet instanser av *rett*N i korpus, ser vi at over halvparten (58%) av instansene er forsøkt tagget for betydning.

3.4.1 Betydningstagget materiale

Tabell 3.3 angir relativ fordeling i ABT-materialet mellom de fire betydninger som *rett*N ble tilordnet ifølge speilmetoden. Som tabellen viser, ble den store majoriteten (73%) av de betydningstaggede instansene assosiert med betydningspartisjonen representert av CLAIM.

Tabell 3.3: Relativ frekvensfordeling i det betydningstaggede materialet mellom de fire betydningene for rettN ifølge speilmetodens betydningspartisjoner

Betydningspartisjon 1, representert ved COURSE	.04
Betydningspartisjon 2, representert ved COURT	.19
Betydningspartisjon 3, representert ved CLAIM	.73
Betydningspartisjon 4, representert ved DISH	.04

Overvekten av instanser tagget som CLAIM er kanskje ikke overraskende: For det første inneholder betydningspartisjonen representert ved CLAIM det største settet av oversettelseskorrespondanser for *rett*N (6 av 13 oversettelseskorrespondanser). Dette vil i seg selv øke sjansene for å finne en oversettelseskorrespondent fra denne betydningpartisjonen. Videre ser vi også at den andre "rettslige" betydningen, representert ved COURT, er den nest mest frekvent valgte betydningskategorien (19%). Dette indikerer at disse to betydningene i sin alminnelighet er mer frekvent i ENPC enn "matrett"-betydningen, hvilket antagelig henger sammen med at sakprosa (deriblant juridiske tekster) utgjør nesten halvparten av tekstmaterialet i parallellkorpuset.

Den ujevne fordelingen mellom betydningene er noe uheldig med tanke på å bruke det automatisk betydningstaggede materialet som input til et maskinlæringssystem for orddisambiguering. Særlig foreligger der få instanser av de to "matrett"-betydningene representert ved COURSE og DISH. Forventningen burde dermed være at maskinlæringssystemet har godt statistisk grunnlag for å lære hva som karakteriserer konteksten til instanser tagget som CLAIM, mens de øvrige tre betydningene sannsynligvis ikke blir like sikre. Det kan også være interessant å se hvorvidt maskinlæringssystemet ihvertfall greier å skille de to "rettslig relaterte" betydningene (COURT og CLAIM) fra "matrett"-betydningene. Dette vil diskuteres nærmere i neste kapittel (4), hvor det betydningstaggede materialet testes på en maskinlæringsalgoritme. Generelt kan det bemerkes at fordelingen mellom betydningene i ENPC antagelig varierer fra lemma til lemma. Basert kun på ett testlemma er det derfor lite motivert å forsøke å generalisere rundt parallellkorpusets relative fordeling mellom betydningene.

En manuell kontroll av testlemmaets betydningstaggede instanser gir grunnlag for følgende observasjoner. I det store og det hele er resultatet tilfredsstillende. 38 galt taggede forekomster ble identifisert, hvilket prosentvis gir 83,3 % riktige betydningstagger. Tabell 3.4 nedenfor presenterer en kategorisering av typene feil som ble funnet, dernest følger en nærmere analyse av feilene. Det skal presiseres at tallet i rad nummer to kun er kalkulert over faktiske forekomster av *rett*N, mens uønskede instanser (instanser av andre lemmaer enn *rett*N) behandles separat som en egen feilkategori (rad nummer tre). Som vi vil se under, kan imidlertid også uønskede instanser som sådan relateres til manglende ordparallellstilling. De

uønskede instansene behandles likevel separat fordi disse *i tillegg* er relatert til usikker preprosessering av korpuset.

Tabell 3.4: Kategorisering av feilene i det betydningstaggede materialet (målt ved relativ frekvens)

Feilkilde i speilmetoden	.21
Tilfeldig generert betydningstag grunnet manglende ordparallellstilling	.21
Uønskede instanser inkludert i det taggede materialet	.58
Totalt antall feil i ABT-materialet (absolutt verdi)	38/228

Med tanke på spørsmålet om parallellkorpusets og speilmetodens anvendelighet som grunnlag for betydningstagging, viser tabell 3.4 at 21 % av feilene (hvilket tilsvarer 8 galt taggede instanser) bunner i en feilkilde fra speilmetoden. I evalueringen av *rett*Ns betydningspartisjoner (2.4.1.) ble det forklart at engelsk *order* kun ble funnet å korrespondere med *rett*N i "matrett"-betydningen. På grunnlag av *order*s øvrige ("ordens"-relaterte) oversettelseskorrespondanser i ENPC, innordnes *order* likevel betydningspartisjonen representert ved CLAIM. Fordi instansen av *rett*N i (8) korresponderer med *order*, har dermed betydningstaggingen av instansen foregått helt korrekt i følge betydningspartisjonene som foreligger - Vi er bare ikke enig i den resulterende semantiske kategorien.

(8) Når en **rett**|CLAIM var klar til servering, ringte madame med en liten klokke på kjøkkenet, og hennes mann hevet øyebrynene i påtatt irritasjon.

De øvrige sju matodisk relaterte feilene er kanskje ikke er like klart "feil", da spørsmålet i disse tilfellene er om en instans burde tagges som CLAIM eller COURT, altså en av de to "rettslige" betydningene. Som eksempel ble instansen i (9) under metodisk sett korrekt tagget som CLAIM fordi *rett*N her korresponderte med engelsk *law*, som tilhører betydningspartisjonen assosiert med CLAIM. Om vi manuelt skulle tagge (9), ville vi kanskje intuitivt velge COURT. Imidlertid er det likevel ikke helt klart om en instans som (9) egentlig burde klassifiseres som "feil": Intuitivt er det egentlig vanskelig å motivere at den ene "rettslig" relaterte betydningen definitivt skulle være mer korrekt enn den andre. I tabell 3.4 ovenfor er likevel de sju tilfellene av denne typen under tvil regnet inn under feil som en følge av metoden for betydningstagging.

(9) Som Arnholm (s. 216) bemerker, var stillingen for barn utenfor ekteskap i vår eldste **rett**|CLAIM slett ikke dårlig.

Tabell 3.4 viser videre at 21 % av feilene er instanser av *rett*N som blir galt betydningstagget pga. manglende ordparallellstilling av korpus, slik at betydningstagger også kan genereres ut fra et lemma som *ikke* korresponderer med instansen som skal tagges. Frekvensandelen på 21 % tilsvarer åtte av betydningstaggerens galt taggede instanser. Som eksempel er engelsk *entitled/be entitled to* det nærmeste vi kommer en korrespondent til *rett*N i (10), og som diskutert under ekserperingsprinsippene er ikke korrespondansen her klar nok til å registrere den.

(10) *Far har bare **rett** til fødselspenger hvis mor gjenopptar arbeidet, tar utdanning på heltid, hun er innlagt i helseinstitusjon eller er så syk at hun er helt avhengig av hjelp fra far for å ta seg av barnet.*
(SI1)
*The father is only **entitled** to maternity benefit if the mother returns to work, takes a **course** of full-time education, is hospitalised or is so ill that she is completely dependent on the father's help with the care of the child.*

Imidlertid viser (10) også at den engelske setningen tilfeldigvis inneholder substantivet *course* (uthevet med fet skrift); lemmaet som representerer *rett*Ns første betydningspartisjon. Følgelig genererer ABT-metoden (som eneste tag) COURSE, og denne tilordnes *rett*N-forekomsten (11):

(11) Far har bare **rett**|COURSE til fødselspenger hvis mor gjenopptar arbeidet, tar utdanning på heltid, hun er innlagt i helseinstitusjon eller er så syk at hun er helt avhengig av hjelp fra far for å ta seg av barnet.

Med tilgang på et ordlenket korpus, ville imidlertid ABT-metoden kunne luke ut slike feil.

Den tredje kategorien av feil som presentert i tabell 3.3 er såkalte *uønskede instanser*, dvs. instanser av andre lemmaer enn søkelemmaet som feilaktig er inkludert i materialet. Denne typen feil utgjør den klart største feilkilden i det betydningstaggede materialet (58 %, som tilsvarer 22 av de 38 identifiserte feilene).

Slike feil kan forklares ved en kombinasjon av mangelen på ordparallellstilling og de åpne søkekriteriene diskutert i (3.3.2). Systemets søkekriterier tillater som vi husker at en instans av *rett*N er tvetydig lemmatisert og pos-tagget. Dette valget ble gjort for å sikre at flest mulig instanser av substantivet *rett*N kan komme opp til vurdering for betydningstagging. Selv om de åpne søkekriteriene også gjør at systemet får treff på uønskede instanser av tvetydig preprosesserte instanser av *rett*A og verbet *rette*, ble det antatt at disse automatisk forkastes fordi systemet (forhåpentligvis) ikke finner noe lemma i den korresponderende setningen som matcher *rett*Ns betydningspartisjoner. Denne antagelsen holder stort sett stikk: I lys av at systemet får treff på til sammen 250 instanser av *rett*A og *rette*V, må det kunne anses som tilfredsstillende at ikke flere enn 22 uønskede instanser feilaktig ble tagget med substantivisk betydning (Alle disse var instanser av adjektivet *rett*). I (12) under gjengis et eksempel, hvor den uønskede instansen ble semantisk kategorisert som COURT. Betydningstaggen COURT ble generert fordi norsk *hoff* tilfeldigvis korresponderer med det engelske substantivet *court*.

(12) Hunts hoff bød på et rikt utvalg av det forkomne og eksotiske , og hver eneste av dem regnet seg som mystiske, uoppdagede talenter, hemmelige stjerner som Hunt var den eneste som hadde erkjent, som Hunt alene kunne stille i sitt **rette**|COURT lys.

Man kan spørre seg om de to sistnevnte feilkildene sier noe om det benyttede parallellkorpusets anvendelighet for et praktisk formål som automatisk betydningstagging. Der synes ikke å være noen grunn til å argumentere for at verken mangelen på ordparallellstiling eller svakheter ved preprosesseringen av ENPC *metodisk* sett utgjør et ankepunkt mot å benytte parallellkorpora som kilde for betydningstagging. Snarere må manglende ordparallelltilling og svakheter ved preprosesseringen av det anvendte korpuset i

denne oppgaven heller anses som en praktisk svakhet spesifikk for ENPC. Det er dessuten et empirisk spørsmål hvorvidt testlemmaet selv utgjør en form for særtilfelle mht. upålitelig lemmatisering/pos-tagging. (For eksempel kan det tenkes at lemmaer på tvers av ordkategorier medfører særlige problemer under den automatiske preprosesseringen.)

Om vi antar at manglende ordparallellstilling og uønskede instanser i det betydningstaggede materialet er å regne som en feilkilde av mer praktisk art (spesifikk for det anvendte parallellkorpuset), kan det som et tankeeksperiment være verdt å se på betydningstaggerens presisjon med hensyn til bare metodiske feil. Om vi tenker oss mer "ideelle" implementeringsmessige omgivelser, som muliggjorde at ABT-metoden luket ut feil av de to sistnevnte typene, gjenstår 198 betydningstaggede instanser hvorav 8 ble galt kategorisert som en følge av en feilkilde i selve metoden for automatisk betydningstagging. I så tilfelle gir metoden med å anvende parallellkorpus og speilmetoden for betydningstagging en presisjon på 96 %. Dette kan kanskje sies å illustrere utslaget det kan gi for resultatene til en applikasjon (som den automatiske betydningstaggeren) hvis den benyttede kilden som gjøres tilgjengelig for eksperimentering ikke selv er optimal og korrekt.

Ordparallellstilling av ENPC er under utvikling i tilknytning til ordnett-prosjektet (jf. 2.3.1). Sannsynligvis kan man dermed forvente at denne feilkilden ikke behøver å utgjøre et problem i samme grad på lang sikt. Innenfor denne oppgaven ble det likevel utført et enkelt forsøk som anvender informasjon om setningsposisjon for å avgjøre om en betydningstag skal benyttes eller om instansen skal forkastes. (Implementeringen foreligger i appendiks 1.) Som tidligere forklart anvender systemet relativ setningsposisjon som et heuristisk mål for å velge mellom flere genererte betydningstagger for en gitt instans. Tanken er da at den betydningstaggen som ble generert fra "nærmeste" posisjon relativt til instansen av *rett*Ns posisjon sannsynligvis er den riktige.

I tråd med en generell antagelse om at et korresponderende setningspar sannsynligvis har noenlunde sammenfallende setningsstruktur, ble det undersøkt om målet på relativ avstand mellom instansen *i* som skal tagges og betydningstaggen *b* også kan brukes for å luke bort enhver vilkårlig generert betydningstag: Når systemet har funnet frem til en betydningstag *b* (evt. utvalgt blant flere mulige), ble det benyttet en terskel for hvor stor den relative avstanden mellom *i* og *b* kan være for at *b* skal tilknyttes *i*. La *n* være instansen *i*s relative setningsposisjon og *m* være setningsposisjonen til betydningstaggen *b*. Hvis den absolutte differansen mellom *n* og *m* er mindre enn en definert terskel, antar vi at *b* ble generert fra en setningsposisjon som er tilstrekkelig nær til å si at *b* sannsynligvis indirekte korresponderer med instansen *i*. Forsøksvis ble målene 1/4 og 1/3 benyttet som terskler. Resultatene for hver terskel ble som følger.

Med krav om at betydningstaggen *b* må ha blitt generert fra en posisjon nærmere enn 1/4 relativt til instansen *i*s respektive setningsposisjon, ble betydningstaggerens presisjon økt til 92,1% (mot opprinnelig 83,3 %). Gitt en terskel på 1/4 ble 24 av de opprinnelige 228 betydningstaggede instansene forkastet fordi betydningstaggen ikke var nær nok. Mht. til de opprinnelige 22 *uønskede* instanser i materialet, ble nå 8 stk. forkastet. Likeledes forkastet systemet nå tre instanser av *rett*N som feilaktig ble tagget som COURSE fordi den korresponderende setningen inneholder uttrykket *of course*. Til gjengjeld gjør terskelen at også 13 setninger som opprinnelig var *korrekt* tagget for betydning blir forkastet.

Hvis vi løser litt på "nærhets"-kravet og setter terskelen til 1/3, reduseres presisjonen til 91,5%, hvilket likevel er bedre enn resultatene slik betydningstaggeren har blitt presentert i denne oppgaven. Med denne terskelen forkastes 16 av de opprinnelige 228 instansene. Av disse ble 9 "korrekt" forkastet (7 adjektiver og 2 av de galt taggede instansene av *rett*N), mens terskelen også gjør at 7 instanser av *rett*N forkastes selv om de i det opprinnelige materialet ble korrekt tagget.

Det er åpenbart at en slik "nærhets"-terskel er temmelig rudimentær sammenlignet med en mer gjennomført ordparallellstilling, og dens fordeler og ulemper må veies mot hverandre. Det fremgår av resultatene at om vi ønsker å øke betydningstaggerens presisjon, så kan en slik terskel benyttes for å forkaste enkelte av de uønskede og galt taggede instansene. På den andre siden blir også opprinnelig korrekt taggede instanser feilaktig forkastet, noe som kanskje er uheldig i et allerede sparsommelig materiale. I denne presentasjonen ble det valgt å la det siste hensynet veie tyngst, og vi forholder oss derfor til betydningstaggerens resultater uten bruk av en "nærhets"-terskel. Dette valget ble gjort fordi det ikke på forhånd er gitt hva som vil utgjøre det største problemet når det betydningstaggede materialet skal benyttes som treningsdata for en maksinlæringsalgoritme. Det er et åpent spørsmål hvorvidt "støy" i treningsdata kompenseres av informasjonen fra noen flere (korrekte) treningsinstanser, eller om et allerede sparsommelig treningskorpus heller bør har en noe høyere presisjon på bekostning av å miste enkelte korrekt taggede instanser.

Neste kapittel går tilbake til ABT-metodens resultat uten bruk av en terskel for "nærhet" mellom en instans og dens mulige betydningstag.

3.4.2 Forkastede (ikke betydningstaggede) instanser

Dette kapittelet gir en kort analyse av instansene av testlemmaet *rett*N som den automatiske betydningstaggeren forkastet. Som vi husker fra ekserperingsprinsippene (2.3.2) ble det slått fast at vi kun registrerer en korrespondent for et søkeuttrykk *x* hvis korrespondenten er klart utskillbar og er tildelt samme type semantiske rolle som *x*. Siden ikke alle instanser av søkelemmaet nødvendigvis har en korrespondent i henhold til ekserperingsprinsippene, er det derfor som forventet at enkelte instanser av *rett*N forkastes av systemet. Det interessante spørsmålet er først og fremst *hvor mange* instanser som ikke kan betydningstagges automatisk med parallellkorpus og speilmetoden som kilde. Et naturlig videre spørsmål er da om ekserperingsprinsippene evt. har vært for strenge.

For ordens skyld presenteres først hvordan de forkastede instansene fordeler seg mellom faktiske instanser av testlemmaet og uønskede forekomster, basert på den automatiske betydningstaggerens data. Tabell 3.5 viser at over halvparten (54%) av de forkastede instansene var hva vi har definert som uønskede instanser. (Disse fordeler seg som 224 tilfeller av *rett*A og 4 forekomster av verbet *rette*.)

Tabell 3.5: Relativ frekvensfordeling mellom forkastede (faktiske) instanser av *rett*N og forkastede uønskede instanser som betydningstaggeren fant i ENPC.

Instanser av *rett*N	.45
Uønskede instanser	.54
Totalt antall forkastede instanser	418

Hva de forkastede faktiske instansene av *rett*N angår, tilsvarer prosentfordelingen på 45% i tabellen 190 instanser. Sammenlignet mot antallet som ble betydningstagget, som beløper seg til 228 instanser[19], ser vi altså at med *rett*N som testlemma tagges drøyt halvparten av instansene mens resten (190) lukes bort.

[19] Med "betydningstagget materiale" menes her det totale antallet, dvs. inkludert de 22 *uønskede* instansene som også kom med.

Under følger noen illustrerende eksempler på instanser som er forkastet fordi instansens korrespondanse ikke er registrert som en gyldig oversettelseskorrespondent til *rett*N i henhold til ekserperingsprinsippene. Instansen i setningen (13) under korresponderer med *entitled/be entitled to*, som ifølge ekserperingsprinsippene ikke regnes som en tilstrekkelig klar oversettelseskorrespondent til *rett*N.

(13) Dette er et land der alle har **rett** til fire ukers betalt ferie, og fem uker der fagforeningene har klart å forhandle seg frem til det.

Korresponderende setning:
This is a nation where everyone is entitled to four weeks paid vacation (five weeks in some positions where individual unions have negotiated five).

(14) og (15) illustrerer at det av og til er vanskelig å finnet et alternativ for hva som overhodet kunne blitt oppført som en utskillbar korrespondanse.

(14) Hvis han bare kunne finne en annen regissør som François Masson, da ville arbeidet hans igjen komme til sin **rett**.

Korresponderende setning:
If only he could find another director like Masson, his work would come into its own again.

(15) En ansatt som må til legen eller tannlegen har **rett** til å gjøre dette på arbeidstid uten trekk i lønna og uten at noen har **rett** til å stille spørsmål.

Korresponderende setning:
A worker with a doctor's or dentist's appointment gets paid time off with no questions asked and the paid time off usually extends also to appointments with physical therapists.

Spørsmålet er først og fremst hva de forkastede instansene forteller oss om ekserperingsprinsippene som ble fulgt, og om det å benytte oversettelseskorrespondanser i et parallellkorpus som ressurs for automatisk betydningstagging. I de tre eksemplene (15-17) ovenfor synes det klart at betydningen til *rett*N manuelt kunne blitt identifisert, i den forstand at selve instansens kontekstuelle betydning ikke er uklar. Oppgavens presenterte ABT-metode kan likevel ikke behandle disse setningene, fordi den er avhengig av at der foreligger en "tilstrekkelig klar korrespondanse" i det motstående språket.

Det ble påpekt i (1.4.4) at man med fordel kan skille mellom to typer orddisambigueringsoppgaver innenfor WSD: Automatisk orddisambiguering som sikter mot å tagge potensielt infinitte mengder av instanser (det "overordnede" målet med WSD), og metoder som prinsipielt sikter mot å ekstrahere et finitt sett av betydningstaggede instanser (til bruk som bakenforliggende ressurs for korpusbasert, overvåket WSD). Siden denne oppgavens ABT-metode legger sistnevnte forutsetning til grunn, er det ikke i seg selv en graverende svakhet at metoden ikke greier å betydningstagge alle ENPCs instanser av testlemmaet.

Imidlertid ønsker vi selvsagt at metoden skal kunne ekstrahere et størst mulig finitt, betydningstagget korpus. Det er derfor naturlig å spørre seg hvorvidt det er "for strenge" ekserperingsprinsipper eller selve parallellkorpusets natur som gjør at ikke flere instanser kan betydningstagges på basis av oversettelseskorrespondanser sortert ved speilmetoden.

Det er vanskelig å forsvare en avgjørelse om å inkludere et eksempel som *entitled* i (13) ovenfor i ekserperingsmaterialet. Grunnen til dette er at det er ønskelig å definere ekserperingsprinsipper som er mest mulig konsekvente (allmenngyldige), samtidig som disse retningslinjene ikke gjør at vi må registrere stikk i strid med hva intuisjonen forteller oss.

Entitled er et typisk eksempel på en korrespondanse som kunne synes både mulig og ønskelig å registrere idet man registrerer korrespondanser for *rett*N. Men når *entitled* i sin tur er søkeuttrykket som vi skal registrere korrespondanser for, synes det straks mer unaturlig å definere hva som skal regnes som en korrespondanse: Skal man da f.eks. bare registrere korrespondanser for partisipp-formen, eller burde man søke på lemmaet *entitle*? (Slik at man m.a.o. registrerer at lemmaet *entitle* korresponderer med *rett*?) Med hensyn til eksemplene (14) og (15) ovenfor, er det av og til heller ikke mulig å finne en god kandidat som oversettelseskorrespondent.

Problemet synes følgelig å være at grunnet forskjeller mellom språk og ikke minst uforutsigbarhet i oversetteren sine valg, så kan man ikke forvente at enhver forekomst av et søkeuttrykk har en definerbar korrespondent i det motstående språket. Basert på resultatene med *rett*N som testlemma, kan i så måte ABT-metoden karakteriseres som et redskap for å ekstrahere et *finitt* sett av betydningstaggede instanser, men som ikke prinsipielt forventes å betydningstagge *alle* instanser som finnes i parallellkorpuset.

3.5 Oppsummering og konklusjon

Dette kapittelet har presentert en tilnærming for automatisk ekstrahering av et finitt, betydningstagget treningskorpus, basert på informasjon fra et parallellkorpus og speilmetoden. Formålet med forsøket var å undersøke i hvilken grad oversettelseskorrespondanser i parallellkorpus, sortert etter betydning ved speilmetoden, er egnet som ressurser for å automatisere betydningstaggingen av et treningskorpus som bakenforliggende ressurs for overvåket WSD. Det har blitt påpekt at en metode for betydningstagging basert på oversettelseskorrespondanser i sin natur er begrenset til de instanser som hadde en identifiserbar, parallellstilt oversettelse i det motstående språket. Med andre ord kan vi si at en slik metode for ABT prinsipielt forventes å betydningstagge et *finitt* korpus, uten å sikte mot å kunne tagge enhver forekomst av det aktuelle flertydige ordet.

Basert på kun ett testlemma har denne oppgavens presenterte resultater et begrenset generaliseringspotensial. Vi observerer likevel at gitt testlemmaet *rett*N, greide ABT-metoden å betydningstagge en stor andel av lemmaets instanser i korpus. (Over halvparten, med *rett*N som testlemma.) Med hensyn til presisjon er 83,3 % av de taggede instansene korrekt (økt til 92,1% ved bruk av en terskel for tillatt avstand mellom instansen og betydningstaggen mht. relativ setningsposisjon.). De galt taggede instansene er stort sett funnet å korrelere med usikker presprosessering av ENPC og av mangelen på ordparallellstilling. Relativt få instanser ble galt betydningstagget som en følge av en feilkilde fra selve speilmetoden. Dette indikerer at den ABT-metoden metodisk sett synes lovende.

Det kan bemerkes at både speilmetodens resultater og selve betydningstaggeren bærer preg av parallellkorpusets størrelse. Siden ENPC per i dag tross alt er relativt lite, har vi sett i evalueringen av speilmetodens resultater at den tvinges til å generere flere betydningsskiller enn hva som kanskje er ønskelig. Av dette følger også for den automatiske semantiske taggingen av et lemma at der blir færre eksempler på hver betydning. Dette må imidlertid kunne forventes å være momenter som kan forandre seg etter som korpusressursene stadig utvikles.

4. Trening av en WSD-klassifikator

4.1 Innledning

Konklusjonen fra forrige kapittel var at det metodisk sett synes lovende å anvende parallellkorpus og speilmetoden som bakenforliggende ressurs for ABT. Det ble også sett at det betydningstaggede materialet inneholder noe "støy" (galt taggede instanser). Disse skyldes i noen tilfeller svakheter ved selve metoden, men primært bunner de i svakheter ved pos-tagging/lemmatisering av det anvendte parallellkorpuset og til en viss grad pga. mangelen på ordparallellstilling i ENPC – Faktorer som er å anse som en praktisk snarere enn en metodisk svakhet.

Et naturlig steg videre er å anvende det ekstraherte betydningstaggede korpuset som treningsmateriale i en overvåket maskinlæringsalgoritme. I overvåket maskinlæring utarbeides en klassifikator for et flertydig ord på grunnlag av et semantisk tagget læringsmateriale som illustrerer korrelasjonen mellom en viss ordbetydning og dens typiske kontekst (jf. 1.4.2). Den vanlige terminologien er å referere til det aktuelle flertydige ordet som *målordet*, altså ordet som skal klassifiseres. Klassifikatoren benyttes så for å klassifisere "nye" instanser av målordet i kontekst (fra nå av: *testinstanser*), det vil si instanser som ikke var del av læringsmaterialet.

Dette kapittelet beskriver treningen av en WSD-klassifikator etter prinsippet for overvåket WSD, som er trent på det betydningstaggede materialet fra den uovervåkede ABT-metoden. Kapittelet er organisert som følger: (4.2) presenterer og motiverer valget av læringsalgoritmen som anvendes i eksperimentene. I (4.3) følger eksperimenter for å kartlegge hvilken type kontekstuell informasjon som synes best egnet for målordet *rett*N, mens (4.4) utgjør den sentrale delen. Her beskrives treningen av tre klassifikatorer for å evaluere det automatisk betydningstaggede materialet.

4.2 TiMBL: Tilburg Memory-Based Learning

Det ble valgt å benytte en minnebasert læringsalgoritme fra softwarepakken TiMBL[20], som er en samling av MBL-algoritmer utviklet ved Tilburg University. (Daelemans et al, 1998).

Innenfor minnebasert læring (MBL), også kalt "lazy learning" (Daelemans et al. 2001), blir alle treningsinstanser lagret i minnet. Klassifiseringen av en testinstans av målordet foregår ved å sammenligne testinstansen mot alle treningsinstanser i minnet, og å klassifisere ut fra den eller de treningsinstanser som ligner mest. Dette prinsippet for å måle likhet kalles *k-nearest neighbour* (*k*-NN), og kan kalkuleres på ulike måter. I (2-3) under vil vi se hvordan *k*-NN beregnes i TiMBL-algoritmen som anvendes i denne oppgaven.

MBLs motstykke er ifølge Daelemans et al. (2001) såkalte "greedy" algoritmer, f.eks. *decision trees* eller *rule induction*. Karakteristisk for en "greedy" algoritme er at den forkaster mindre frekvente kontekstmønster ved å anvende en form for abstraheringsmetode under læringen, for eksempel en frekvensbasert metode: Treningsmaterialet abstraheres til en modell (hypotese) om hva som antas å være mest karakteristisk for hver av kategoriene (f.eks betydningskategoriene) som skal læres.

Daelemans et al. (1999) har vist at for de fleste NLP-oppgaver yter MBL-algoritmer best, fordi systemet da kan klassifisere en testinstans ved å sammenligne likheten/distansen (*k*-NN) mellom testinstansen og *alle* treningsinstanser systemet ble presentert for; selv lavfrekvente eksempler. MBLs *k*-NN likhetsmål innebærer selvsagt ikke at algoritmen er ment å finne en *identisk* treningsinstans; poenget er å finne den/de *nærmeste* "naboene" til testinstansen. Dette betyr for det første at MBL-algoritmer, til forskjell fra "greedy" algoritmer, ikke restrukturerer eller abstraherer bort fra treningsmaterialet. Siden MBL for det andre heller ikke krever helt identisk likhet inneholder de likevel en (ønsket) evne til å generalisere ut fra treningsmaterialet. Gitt at treningsmaterialet i denne oppgaven i utgangspunktet er relativt lite, synes det derfor mest formålstjenlig å anvende en MBL-algoritme, for å beholde flest mulig eksempelinstanser som grunnlag for klassifisering.

Input til en TiMBL-algoritme er et sett av treningsinstanser, som i vårt tilfelle utgjør eksempelforekomster av et flertydig ord (målord) i kontekst. En treningsinstans representeres som en trekkvektor, og hver slik trekkvektor må bestå av en fast mengde av trekkverdier (kontekstord) og riktig kategori (riktig betydningstag). I eksempelet (1) under gjengis en treningsinstans med fem trekk: Den umiddelbare, lokale konteksten rundt testlemmaet *rett*N i et kontekstvindu på [± 2] samt selve målordet i midten. (Strengt tatt behøver ikke selve instansen av *rett*N være med i trekkvektoren, men her tas den med for illustrasjonens del.) Til sist følger riktig betydningstag, det vil si den betydningskategorien som er assosiert med målordet i denne konteksten. Inputformatet som vil anvendes i denne oppgaven er *comma-separated values* (*csv*-format), det vil si at hver trekkverdi er separert med komma.

(1) `beiterett,og,rett,til,brensel,claim`

I denne oppgaven benyttes TiMBLs *default*-algoritme IB1. IB1 kalkulerer *k*-NN ved overlappingsmålet i (2) og (3) under (Daelemans et al, 2001). Algoritmen finner de *k* nærmeste naboene (vanligvis settes *k* til 1) til en testinstans ved å måle *distansen* mellom en

[20] Tilgjengelig fra http://ilk.kub.nl

treningsinstans og testinstans, $\Delta(X, Y)$. Instansene er da representert ved n trekk, og δ er avstanden per trekkverdi. Distansen mellom to instanser regnes ut ved summen av instansenes *ulike* trekkverdier.

(2) $$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

hvor det for ikke-numeriske trekkverdier (som kontekstordene i denne oppgaven) gjelder at:

(3) $$\delta(x_i, y_i) = 0 \quad \text{hvis } x_i = y_i$$
$$= 1 \quad \text{hvis } x_i \neq y_i$$

Dette målet kan videre kombineres med andre mål som påvirker definisjonen på likhet, for eksempel ved å vekte hvert trekks relevans. Hvis det blir funnet flere enn én "nærmeste nabo" med samme likhetsdistanse i forhold til testinstansen, velges majoritets-kategorien blant de funnede "naboene". (Hvis klassifikatoren finner tre naboer, hvor to eksemplifiserer CLAIM og den siste COURT, klassifiseres testinstansen som CLAIM.) Hvis der er like mange "naboer" av hver kategori, for eksempel to CLAIM og to COURT, selekteres kategorien med høyest total frekvens i hele treningsmaterialet (hvilket i vårt eksempel ville være CLAIM).

Vi vil utføre både trening og testing av en klassifikator på det samme treningsmaterialet: Ved testmetoden kryssvalidering splittes treningsmaterialet inn i n partisjoner (f.eks. $n = 10$) slik at hver partisjon n_i etter tur benyttes som testmateriale, med de resterende 1-n_i som treningsmateriale. Framfor å trene én stor klassifikator på det totale treningsmaterialet (som deretter testes på et eksternt testkorpus), innbærer altså kryssvalidering at det trenes n (f.eks. $n = 10$) delklassifikatorer ved adskilte tester, én per partisjon. Den engelske termen er "n-fold cross-validation", for eksempel "10-fold cross validation". På norsk vil vi fra nå av bruke termen "n-delt kryssvalidering", for eksempel tidelt kryssvalidering.

4.3 Kartlegging av kontekstuell informasjon for testlemmaet

4.3.1 Innledning

Siden treningseksempler kodes i TiMBLs input som trekkvektorer med et fast antall trekk, må altså alle treningsinstanser av målordet ha et konstant antall *n* trekkverdier (kontekstord) i sin trekkvektor. (Antallet i et eksperiment velges av brukeren, men må være likt for alle instanser som inngår i eksperimentet.)

Første steg for å trene en klassifikator var derfor å avgjøre hvilken kontekstype og størrelse på kontekstvinduet som skal ekstraheres fra korpuset for hver ordforekomst. Dette spørsmålet kunne i seg selv utgjøre en hel oppgave, da der er mange muligheter for hvilken lingvistisk informasjon som kan bidra: Skal man bruke lemmatisert kontekst eller bøyde former; bør man inkludere pos-informasjon for hvert kontekstord; og skal man kun basere seg på den foreliggende konteksten, eller i tillegg inkludere semantisk informasjon fra en ekstern kunnskapsressurs (f.eks. informasjon om semantisk relaterte ord i følge et tesaurus eller et ordnett)? (Resnik & Yarowsky, 1997/2000; Leacock & Chodorow, 1998; Stevenson & Wilks, 2001, m.fl).

Innenfor rammene av denne oppgaven ble det valgt å basere seg på kun konteksten som foreligger ifølge korpus (uten eksterne informasjonskilder). Siden treningsmaterialet er relativt lite, vil det være en fordel å hente ut kontekst i lemmatisert form, slik at konteksten normaliseres med hensyn til variasjon i bøyning (og følgelig supplerer flere sammenlignbare eksempler). Videre har vi sett at pos-taggingen i ENPC ikke er manuelt korrigert, og kan dermed utgjøre en feilkilde. Derfor vil ikke informasjon om pos-tagger tilknyttes hvert ord.

Spørsmålet som da står igjen er hvilken type kontekst og størrelse på kontekstvindu som skal ekstraheres fra korpuset. Chodorow et al. (1998) og Leacock & Chodorow (1998) har gjennomført systematiske eksperimenter på bruken av *lokal, posisjonsspesifikk kontekst* rundt det aktuelle ordet versus bruk av *nøkkelord* (åpen klasse-ord i et større vindu), samt en kombinasjon av disse. Deres eksperimenter antyder at for substantiver (som er relevant for denne oppgaven) synes nøkkelord å være noe mer informativt enn lokal kontekst; men aller best fungerer en kombinasjon av begge typer informasjon. Imidlertid har det blitt påpekt av flere (Resnik & Yarowsky, 2000; Stevenson & Wilks, 2001; Hoste et al., 1998) at orddisambiguering er en oppgave som later til å være høyst leksikalsk sensitiv, slik at en klassifikators performans kan variere fra testlemma til testlemma. Det synes derfor formålstjenlig å kartlegge egenskapene til denne oppgavens testlemma, *rett*N, med hensyn til hvilken kontekstuell informasjon som fungerer best. Denne "beste" typen informasjon vil bli anvendt i evalueringseksperimentene av ABT-materialet i (4.4).

Dette delkapittelet presenterer noen enkle eksperimenter for å undersøke (I) optimal størrelse på lokalt kontekstvindu, (II) kontekstuelle nøkkelord, og (III) en kombinasjon av begge typer kontekst. (For LISP-implementering av hvordan de aktuelle kontekstypene ekstraheres fra korpus, se appendiks (2).) Forsøkene ble utført ved TiMBLs kryssvalideringsmetode, hvor treningkorpuset ble oppdelt i 10 partisjoner. Eksperimentene for å kartlegge kontekstuell informasjon vil ikke vies en detaljert analyse av hva som slår feil og hvorfor; dette vil gjøres under evalueringen av ABT-materialet i (4.4).

4.3.2 Eksperiment 1: Lokal, posisjonsspesifikk kontekst

Lokal, posisjonsspesifikk kontekst består av et kontekstvindu på [± *n*], som inkluderer de *n* tegnene umiddelbart før og etter det flertydige ordet. Ifølge Leacock & Chodorow (1998) har et lokalt kontekstvindu på [± 2] blitt funnet å fungere best, også i forbindelse med andres tidligere eksperimenter[21]. I denne oppgaven verifiserer vi dette ved å trene tre lokalkontekst-klassifikatorer på det automatisk betydningstaggede materialet, hvor størrelsen på kontekstvinduet er henholdsvis [± 2], [± 4] og [± 6]. Eksperimentene utføres ved tidelt kryssvalidering. M.a.o. deles treningskorpuset i ti partisjoner, og innenfor ett eksperiment trenes og testes ti adskilte "del-klassifikatorer".

I denne oppgavens implementering (appendiks (2)) ble lokal kontekst ekstrahert som følger. Det er vanlig å telle ikke bare ord men også tegnsetting i et lokalt kontekstvindu. Om det flertydige ordet (målordet) f.eks. følges av et punktum eller en tankestrek, tas altså dette tegnet med som en egen trekkverdi. Videre er vi intuitivt kun interessert i lokal kontekst som inngår i samme setning som målordet. Grunnen til dette er at det/de siste ordene i den forrige setningen eller det/de første ordene i neste setning neppe vil være spesielt karakteristiske for målordets betydning innenfor sin gitte setning. Følgelig, hvis der ikke er tilstrekkelig mange kontekstord før eller etter målordet innenfor målordets setning, fylles de "tomme" posisjonene i trekkvektoren med et bindestrek; "-". I eksempel (4) under, med et kontekstvindu på [± 2], var testlemmaet *rett*N første ord i setningen. Følgelig fylles de to posisjonene foran ordet med "-".

(4) -,-,rett,gjelde,både,claim

Tabell 4.1 gjengir resultatene av de tre lokalkontekst-klassifikatorene, testet ved tidelt kryssvalidering på det automatisk betydningstaggede treningsmaterialet med TiMBLs læringsalgoritme IB1. Tabellen gjengir antall korrekte tagger for hvert av kryssvalideringsmetodens ti delresultater, samt gjennomsnittlig prosentvis presisjon for hvert kontekstvindus klassifikator. (I dette delkapittelet vil vi ikke gå inn på hvordan klassifiseringsresultatene fordeler seg mellom de fire betydningene av *rett*N; dette vil vi komme tilbake til i (4.4).

[21] Blant andre nevner de Hearst (1991), Resnik (1993b), Yarowsky (1993), Leacock et al (1996) og Bruce & Wiebe (1994).

Tabell 4.1: Maskinlæringsresultater ved tidelt kryssvalidering på automatisk betydningstagget treningsmateriale: Lokalt kontekstvindu.

Kontekstvindu	**[± 2]**	**[± 4]**	**[± 6]**
Delresultater (antall korrekte av totalt antall instanser i hvert testsett)	21/23	19/23	19/23
	21/23	20/23	20/23
	18/23	18/23	19/23
	13/23	14/23	15/23
	19/23	17/23	17/23
	7/23	6/23	7/23
	18/23	17/23	18/23
	20/23	18/23	16/23
	21/23	21/23	21/23
	19/21	19/21	19/21
Totalt antall riktige	177/228	169/228	171/228
Gj.snittlig presisjon %	**77,63 %**	**74,12 %**	**75 %**

Resultatene i tabell 4.1 viser at et lokalt kontekstvindu på [± 2] gjennomsnittlig klassifiserer best (77,63 %). Det automatisk betydningstaggede korpuset ble partisjonert likt for alle tre kontekstvinduer, og for de fleste delresultatene ser vi en klar sammenheng mellom alle tre kontekstvinduers presisjon for et gitt testsett. For eksempel blir det sjette testsettet svakest klassifisert ved alle tre kontekstvinduer. En nærmere analyse viste at dette settet tilfeldigvis inneholder hele 21 av de totalt 43 instansene som er tagget som COURT i det totale betydningstaggede treningsmaterialet. Når denne partisjonen plukkes ut som testsett (mens resten av korpuset utgjør treningsmateriale), mister klassifikatoren følgelig halve treningsgrunnlaget for å gjenkjenne mulige kontekster assosiert med COURT. En mulig løsning på dette kunne være å randomisere treningsmaterialet før det partisjoneres, dette ble imidlertid ikke gjort innenfor rammene av denne oppgaven.

Åpenbart er det begrenset hva man kan konkludere ut fra testing på bare ett testlemma, men vi ser altså at testlemmaet *rett*N ikke later til å motsi tidligere eksperimenters observasjon om at et kontekstvindu på [± 2] synes å være en pålitelig kilde for orddisambiguering (Leacock & Chodorow, 1998).

4.3.3 Eksperiment 2: Nøkkelord i et større kontekstvindu

Formålet med å hente ut såkalte nøkkelord fra en større kontekst er å ekstrahere informasjon om hvilket tematisk domene den flertydige ordforekomsten inngår i. Nøkkelord kan defineres etter mange kriterier. I Leacock & Chodorows (1998) eksperiment defineres nøkkelord som ord innenfor et større vindu på [± *n*] som *ikke* er grammatiske ord; det vil si substantiver, verb, adjektiver og adverb. (Intuitivt vil grammatiske ord neppe være særlig informative for å antyde tema, siden de generelt kan inngå i enhver kontekst.)

Innenfor denne oppgaven ble det implementert to ulike definisjoner for å hente ut nøkkelord. Felles for begge er at vi definerer et kontekstvindu på [*n*] som henter ut de *n* nærmeste nøkkelordene for hver instans av målordet *rett*N, innenfor instansens dokument. Dette innebærer å begynne ved den aktuelle instansen, og å søke etter de første *n*/2 ord på begge sider av ordforekomsten som tilfredsstiller betingelsene for å regnes som et nøkkelord. I utgangspunktet foregår altså søket likt på begge sider av målordet, etter [± *n*]-mønsteret vi så

for lokal kontekst. Imidlertid, siden TiMBL krever at alle trekkvektorer må inneholde like mange trekk, var det nødvendig å ta høyde for at der kanskje ikke er tilstrekkelig mange nøkkelord på hver side. (For eksempel hvis målordet forekommer blant det aktuelle dokumentets første eller siste setninger.) Derfor defineres heller et vindu [*n*] som teller opp like mange ord ([± *n*/2]) på hver side av ordforekomsten hvis mulig, men som ellers utvider tellingen på den kontekstsiden hvor der finnes nok ord.

De to ulike nøkkelorddefinisjonene som ble testet ut er som følger.

(I) Nøkkelord I: [*n*] nærmeste ikke-grammatiske og ikke-frekvente ord

Denne definisjonen, som vi vil referere til som *nøkkelord I*, bruker to kriterier for å hente ut nøkkelord: For det første anvendes ENPCs pos-tagger for å luke bort grammatiske ord. Med forbehold om feil i preprosesseringen av ENPC, inkluderes altså kun substantiver, verb, adjektiver og adverb (åpen klasse-ord) blant nøkkelordene. For det andre kan det i tillegg være ønskelig å luke ut hjelpeverb (modaler og hjelpeverbet *ha*) og andre åpen klasse-ord som generelt er svært frekvent i norsk (f.eks. adverber som *imidlertid* og *ofte*). Hvis et åpen klasse-ord generelt er svært frekvent i norsk, vil det sannsynligvis være lite informativt for å antyde et bestemt tema. Derfor luker vi ord som er oppført i `Tekstlaboratoriets liste over de 10.000 mest frekvente norske ord, basert på Oslo-korpuset av taggede, norske tekster`[22].

(II) Nøkkelord II: [*n*] nærmeste karakteristiske nøkkelord

Det ble som et alternativ forsøkt å definere de *n* nærmeste, mest *karakteristiske* nøkkelord i et flertydig ords omgivelser (*nøkkelord II*). "Karakteristisk" defineres ved å måle kontekstordets relative frekvens innenfor det aktuelle dokumentet mot kontekstordets relative frekvens i hele ENPC-korpuset. Intuisjonen er at hvis kontekstordets relative frekvens i det gitte dokumentet er høyere enn dets relative frekvens i hele ENPC, så kan vi anta at ordet er karakteristisk for dette dokumentet. Et ord *x* defineres således som "karakteristisk" for det aktuelle dokumentet *d*, etter følgende mål, gjengitt i (5):

$$(5) \qquad \frac{frekvens(x \mid d)}{antall\ ord(d)} \rangle \frac{frekvens(x \mid ENPC)}{antall\ ord(ENPC)}$$

Tanken bak denne definisjonen var opprinnelig å hente ut ethvert *n* nærmeste ord som oppfyller "karakteristisk"-kriteriet i (5); uten hensyn til om ordet er av lukket eller åpen ordklasse. Det viste seg imidlertid at det da blir inkludert flere ord fra lukkede ordklasser enn forventet, og at disse vanskelig kan regnes som *de facto* karakteristiske nøkkelord for en gitt ordbetydning (f.eks. ord som *ikke*, *mye* og *mange*). Derfor ble det valgt å luke ut lukkede ordklasser også i denne definisjonen, for å hente ut de mer semantisk informative *n* nærmeste karakteristiske nøkkelord av åpen klasse.

Under gjengis kontekstordene som ble ekstrahert ved definisjonene av henholdsvis nøkkelord I og nøkkelord II for de samme instanser av betydningene CLAIM (6) og COURSE (7). Ordene som er felles for henholdsvis nøkkelord I og nøkkelord II er uthevet med fet skrift.

[22] Frekvensordlisten er basert på Oslokorpuset, som inneholder 18,3 millioner ord fra tekstgenrene skjønnlitteratur, avis/ukeblad og sakprosa. Tekstlaboratoriets nettside er: http://www.hf.uio.no/tekstlab/.

Kontekstvindu: [30]

(6) Nøkkelord I:
råd,**fylkesting**,**kommunestyre**,**fylkesråd**,**Kommunerådets**,**ledelsesansvar**,prinsipiell, enkeltsak,enkeltmedlem,avgjørelsesmyndighet,**tildele**,Kommunestyret,påse,Kommunerådet ,**fylkeskommunal**,**kontrollutvalg**,ha,**ansvarsområde**,**Kontormessig**,**utredningsmessig**, **assistanse**,**kommunestyremedlem**,fylkestingsmedlem,kommuneråd,kontormessig,kapittel, Kommunens,fylkeskommune,Administrasjonssjef,administrasjonssjef,claim

Nøkkelord II:
råd,møte,**fylkesting**,**kommunestyre**,delta,skulle,medlem,**fylkesråd**,**Kommunerådets**, bestemme,nr.,**ledelsesansvar**,**tildele**,tilfelle,betydning,kommunal,**fylkeskommunal**,organ, unntak,**kontrollutvalg**,personlig,øvrig,enkelt,rett,folkevalgt,**ansvarsområde**,**Kontormessig**, **utredningsmessig**,**assistanse**,**kommunestyremedlem**,claim

(7) Nøkkelord I:
omløpstid,**innrette**,**sted**,**Ami**,**Bel**,effektuere,bestilling,duk,rutemønster,gaffel,dessert,frukt kompott,gryte,indonesisk,lapskaus,**prefabrikat**,**dypfrossen**,**delikatesse**,**pommes_frites**, **ingrediens**,**frityrolje**,blid,kresen,matvei,serveringsdame,egentlig,vanke,medgi,smil, sympatisk,course

Nøkkelord II:
omløpstid,rask,**innrette**,alt,meny,gå,spise,slik,**sted**,**Ami**,**Bel**,tid,aldri,ta,mat,basere, **prefabrikat**,**dypfrossen**,**delikatesse**,**pommes_frites**,egentlig,fersk,**ingrediens**,se,**frityrolje**, bære,smak,gårsdag,dag,servere,course

Eksemplene ovenfor antyder at ca. halvparten av nøkkelordene blir felles for de to definisjonene. Ut fra øvrige manuelle stikkprøver i treningsmaterialet kan vi grovt estimere at mellom 25-50 % av kontekstordene blir felles. Generelt gjelder at definisjon I later til å være strengere enn definisjon II, som vi også kan se i (6) og (7) ovenfor. I (7) følger for eksempel *innrette* og *sted* direkte etter hverandre for nøkkelord I, mens nøkkelord II har inkludert ytterligere fem kontekstord mellom disse to ordene (*alt, meny, gå*, *spise* og *slik*). Dette innbærer med andre ord at nøkkelord I-definisjonen generelt ekstraherer kontekstord fra et større tekstområde (det vil si fjernere fra målordet). Videre antyder (6) og (7) en tendens til at de karakteristiske nøkkelord II, i noe større grad enn nøkkelord I, inkluderer kontekstord som vanskelig kan sies å være relevante for en spesifikk ordbetydning (f.eks. *skulle* i (6) og *alt, gå, slik, aldri, ta* og *se* i (7)). Imidlertid gjelder dette også til en viss grad for nøkkelord I, for eksempel ord som *ha* i (6) og *blid* og *sympatisk* i (7).

I eksperimentet ble det valgt å teste tre ulike størrelser på kontekstvinduet for hver av de to definisjonene: [12], [30] og [50]. Testingen ble utført på samme måte som for lokal kontekst, det vil si ved tidelt kryssvalidering på hvert av de tre kontekstvinduene for henholdsvis nøkkelord I og (karakteristiske) nøkkelord II. Resultatene presenteres i følgende tabell 4.2.

Tabell 4.2: Maskinlæringsresultater ved tidelt kryssvalidering på automatisk betydningstagget treningsmateriale: *Nøkkelord I* og (*karakteristiske) nøkkelord* II.

Kontekstvindu	**[12]**		**[30]**		**[50]**	
Type nøkkelord	Nøkkelord I	Nøkkelord II	Nøkkelord I	Nøkkelord II	Nøkkelord I	Nøkkelord II
Delresultater (antall korrekte av totalt antall instanser i hvert testsett)	19/23	18/23	17/23	19/23	17/23	18/23
	19/23	19/23	17/23	18/23	14/23	20/23
	15/23	16/23	16/23	12/23	14/23	17/23
	14/23	13/23	15/23	12/23	15/23	14/23
	16/23	15/23	14/23	10/23	14/23	14/23
	4/23	9/23	4/23	5/23	1/23	6/23
	19/23	17/23	17/23	15/23	16/23	17/23
	15/23	13/23	15/23	12/23	15/23	15/23
	21/23	17/23	21/23	20/23	20/23	18/23
	15/21	19/21	15/21	15/21	14/21	17/21
Totalt antall riktige	157/228	156/228	151/228	138/228	140/228	156/228
Gj.snittlig presisjon %	**68,86 %**	**68,42 %**	**66,23 %**	**60,53 %**	**61,40 %**	**68,42 %**

Resultatene i tabell 4.2 er interessante, fordi de først og fremst viser hvor vanskelig det kan være å avgjøre både vindusstørrelse og hvilken type nøkkelord som fungerer best. Kort sagt synes forskjellene både i delresultater og gjennomsnitt så lite systematiske at det er vanskelig å peke ut en klar tendens. Sannsynligvis bunner dette primært i det totalt sett sparsomme statistiske grunnlaget, og dessuten kan også tilfeldigheter ved partisjoneringen av treningskorpuset ha spilt en rolle.

Med hensyn til vindusstørrelse observerer vi at nøkkelordvinduet [12] gjennomsnittlig klassifiserer best, med 68,86 % for nøkkelord av definisjon I. Imidlertid er ikke dette beste resultatet signifikant bedre enn alle øvrige resultater: Faktisk har nøkkelord II-definisjonens vindu på [50] et nesten like høyt nivå, 68,42 %. På den andre siden ser vi at kontekstvinduet på [50] for nøkkelord av definisjon I presterer nest svakest av alle seks klassifikatorer, med et gjennomsnitt på 61,40 %. Hva de to definisjonene av nøkkelord angår, antyder delresultatene og gjennomsnittet til vinduene [12] og [30] en liten tendens til at den første definisjonen (nøkkelord I) yter like godt eller bedre enn (karakteristiske) nøkkelord II. I vinduet på [50] ord er derimot den klare tendensen i delresultatene og gjennomsnitt at nøkkelord II presterer bedre enn nøkkelord I.

Som vi ser, synes det åpenbart at spørsmålet om nøkkelords definisjon og størrelse hadde fortjent en grundigere utforskning. Innenfor rammene av denne oppgaven, hvor vi bare har tilgang på ett testlemma (som i tillegg er supplert med et relativt sparsomt treningsmateriale), synes det imidlertid lite velmotivert å forsøke å utforske spørsmålet om nøkkelord i dybden.

I det følgende vil vi derfor ganske enkelt ta utgangspunkt i de beste resultatene som foreligger for henholdsvis lokal kontekst og nøkkelord, og forsøke å kombinere disse.

4.3.4 Eksperiment 3: En kombinasjon av lokal kontekst og nøkkelord

Som nevnt i innledningen (4.3.1), har Chodorow et al. (1998) og Leacock & Chodorow (1998) funnet at substantiver later til å bli best klassifisert med utgangspunkt i en kombinasjon av lokal kontekst og nøkkelord. Vi vil her undersøke dette ved å trene en klassifikator som bruker kontekstinformasjonen fra de beste resultatene som foreligger fra eksperimentene 1 (4.3.2) og 2 (4.3.3). Det ble således ekstrahert en ny kontekst for det automatisk betydningstaggede treningsmaterialet av *rett*N, som består av:

(I) Lokal, posisjonsspesifikk kontekst i et vindu på [± 2] (som beskrevet i (4.3.2))
(II) De 12 nærmeste nøkkelord etter den første definisjonen i (4.3.3), det vil si hvor systemet luker ut grammatiske ord (på basis av ENPCs pos-tagger) og de 10.000 mest frekvente ord i norsk (ifølge Tekstlaboratoriets frekvensliste).

Denne kontekstypens trekkvektorer består da av 17 trekk[23], og vi kan referere til kontekstypen som Lokal&Nøkkel17. Eksempel (8) under illustrerer den samme eksempelinstansen som (6) i (4.3.3) ovenfor.

(8) NIL,ha,rett,til,å,råd,fylkesting,kommunestyre,fylkesråd,Kommunerådets,ledelsesansvar, fylkeskommunal,kontrollutvalg,ha,ansvarsområde,Kontormessig,utredningsmessig,claim

Eksperimentet ble utført på samme måte som ved lokal kontekst og nøkkelord, det vil si med det betydningstaggede materialet som basis, og utført ved kryssvalidering med TiMBLs IB1-algoritme. Resultatet er presentert i tabell 4.3 under, hvor også resultatene for lokal kontekst på [± 2] og de 12 nærmeste nøkkelordene er inkludert for sammenligningens del.

Tabell 4.3: Maskinlæringsresultater ved tidelt kryssvalidering på automatisk betydningstagget treningsmateriale: Lokal kontekst, nøkkelord og kombinasjonen *Lokal&Nøkkel17*

Kontekstype	**Lokal kontekst [± 2]**	**Nøkkelord [12]**	**Lokal&Nøkkel17**
Delresultater (antall korrekte av totalt antall instanser i testsettet)	21/23	19/23	21/23
	21/23	19/23	22/23
	18/23	15/23	18/23
	13/23	14/23	14/23
	19/23	16/23	18/23
	7/23	4/23	8/23
	18/23	19/23	18/23
	20/23	15/23	20/23
	21/23	21/23	21/23
	19/21	15/21	18/21
Totalt antall korrekte	177/228	157/228	178/228
Gj.snittlig presisjon %	**77,63 %**	**68,86 %**	**78,10 %**

[23] Det 17. trekket kommer av at den lokale konteksten eksplisitt inkluderer den flertydige forekomsten av *rett*N (slik at et kontekstvindu på [± 2] utgjør 5 trekk og ikke 4). Det er strengt tatt ikke nødvendig å ta med dette femte trekket, men da den heller ikke gjør skade (siden dette trekket alltid er tilstede og på samme plass) ble den inkludert for bedre oversikt under den manuelle evalueringen.

Som vi ser, later Lokal&Nøkkel17s resultater til å bekrefte at en kombinasjon av lokal kontekst og nøkkelord fungerer noe bedre enn bare én av kontekstypene. Vi observerer imidlertid at der ikke er en stor forskjell mellom lokal kontekst alene og kombinasjonen av lokal kontekst og nøkkelord: I bare tre av delresultatene presterer Lokal&Nøkkel17 bedre enn lokal kontekst alene, to ganger ytte lokal kontekst bedre, og i de øvrige fem delresultatene hadde lokal kontekst og nøkkelord sammenfallende resultater.

Det skal her nevnes at kombinasjonen av flere typer informasjon kan gjennomføres på flere måter. Innenfor rammene av denne oppgaven er de to typene kontekstinformasjon rett og slett slått sammen for hver trekkvektor. Imidlertid har for eksempel Hoste et al. (2001) forsøkt å trene fire delklassifikatorer for hvert flertydige ord *x*: Klassifikator 1 bruker informasjon om lokal kontekst, nummer 2 bruker nøkkelord, nummer 3 anvender en kombinasjon, mens en "default" nummer 4 baserer seg på den mest frekvente betydningen totalt sett. For hver testinstans av *x* anvendes alle fire klassifikatorer, og hver av dem gir hvert sitt forslag til riktig betydningstag. Det endelige valget av betydningstag utføres ved majoritetsvalg (den betydningstaggen som ble foreslått av flest delklassifikatorer) eller ved vektet valg (hvor "stemmen" til del-klassifikatorene med høyest total presisjon vektes som viktigere enn de mindre pålitelige klassifikatorene).

4.3.5 Oppsummering

Dette delkapittelet hadde som mål å utføre noen enkle eksperimenter for å kartlegge hvilken type kontekst vi vil anvende i neste delkapittels eksperimenter for å evaluere det betydningstaggede materialet. Tre typer kontekst ble ekstrahert for det automatisk betydningstaggede materialet: Lokal, posisjonsspesifikk kontekst, to ulike definisjoner av nøkkelord, og til slutt en kombinasjon av det beste lokale vinduet ([$\pm$ 2]) og nøkkelordvinduet ([12] etter definisjon I i (4.3.3). Resultatene viste at kombinasjonen av lokal kontekst og nøkkelord (Lokal&Nøkkel17) klassifiserer svakt bedre enn lokal kontekst (med en gjennomsnittlig presisjon på 78,10 % mot lokal kontekst 77,63 %).

Imidlertid må det påpekes at disse resultatene sannsynligvis er påvirket av det relativt sparsommelige treningsmaterialet som har inngått i eksperimentene samt av tilfeldigheter ved partisjoneringen av det totale treningsmaterialet. I tillegg gir denne oppgaven resultatene for bare ett testlemmas automatisk betydningstaggede korpus (*rett*Ns). Innenfor rammene av denne oppgaven syntes det derfor lite velmotivert å gå videre i dybden av spørsmålet om hvilken kontekstype som egner seg best for orddisambiguering. Det ville imidlertid vært interessant å undersøke spesielt ekstraheringen av nøkkelord mer i detalj, fortrinnsvis med tilgang på flere testlemmaer. Innenfor denne oppgaven velger vi imidlertid å forholde oss til det beste resultatet som forligger for de videre eksperimentene, nemlig klassifikatoren Lokal&Nøkkel17.

4.4 Trening av tre WSD-klassifikatorer for å evaluere det automatisk betydningstaggede materialet

4.4.1 Innledning

Formålet i dette delkapittelet er å gi en nærmere analyse av ABT-materialets potensial som treningsdata for en WSD-klassifikator. Dette vil gjøres ved å trene tre klassifikatorer, som forklart i det følgende.

Den beste klassifikatoren fra (4.3), Lokal&Nøkkelord17, vil vi fra nå av referere til som klassifikatoren K1. Som vi husker, er K1 trent på ABT-metodens materiale for testlemmaet *rett*N. K1 kan således sies å antyde hvor godt maskinlæringsalgoritmen presterer ved bruk av et automatisk ekstrahert treningskorpus; slik metoden for ABT er implementert i denne oppgaven og med parallellkorpuset ENPC som testressurs.

Vi så i evalueringen av ABT-materialet i (2.5.4) at K1s treningsmateriale både inneholder feil som er knyttet til selve ABT-metoden og feil av mer praktisk art (i sammenheng med mangelen på ordparallellstiling og svak preprosessering av ENPC). For bedre å kunne evaluere potensialet til selve metoden for ABT som er presentert i denne oppgaven, vil vi korrigere ABT-materialet for de feil som er av praktisk snarere enn av metodisk natur. En klassifikator K2 som er trent på dette delvis korrigerte materialet kan således betraktes som en indikator på hvordan selve metoden for ABT egner seg for å supplere treningsmateriale for en WSD-klassifikator, under "ideelle" implementeringsmessige omgivelser for øvrig. Målet med å sammenligne K1 mot K2 er å få et bedre bilde av hvilken rolle metodiske feilkilder spiller i forhold til de mer "praktisk" relaterte feil i treningen av en WSD-klassifikator.

Siden den presenterte metoden for ABT sikter mot å løse problemet med manglende betydningstaggede korpora (hvor taggingen vanligvis må foregå manuelt), er det naturlig å sammenligne klassifikatoren K2 mot en klassifikator som er trent på et manuelt betydningstagget treningskorpus. Det manuelt betydningstaggede korpuset består av alle instanser av testlemmaet *rett*N i ENPC, og har blitt manuelt tagget med speilmetodens betydninger for testlemmaet. Denne klassifikatoren K3 kan betraktes som en "baseline" eller "gullstandard" for sammenligning av ABT-metodens begrensninger i forhold til et manuelt tagget treningskorpus. K3 er ment å indikere hvor langt det var mulig å komme i treningen av en klassifikator, med utgangpunkt i de instanser som ENPC supplerer for *rett*N og med speilmetodens betydningsskiller. Spørsmålet er hvor avgjørende de metodiske feilkildene i K2s treningsmaterialet er for klassifikatorens læringsevne, sett i forhold til K3.

Treningen av de to klassifikatorene K2 og K3 vil foregå på samme måte som for K1, det vil si ved tidelt kryssvalidering på treningsmaterialet med TiMBLs læringsalgoritme IB1. (Slik at treningsmaterialet tjener som både trenings- og testmateriale.)

Analysen av klassifikatorene er lagt opp som følger. Det ble valgt å ikke sikte mot en statistisk analyse. For det første er statistiske analyser forbundet med formelle krav til for eksempel normalfordeling og likt standardavvik mellom gruppene som sammenlignes. Siden det statistiske grunnlaget i hver av de tre klassifikatorenes betydningstaggede korpora er relativt sparsomt, er det vanskelig å oppfylle slike formelle krav for statistisk "sikre" observasjoner. Siden denne oppgaven dessuten konsentrerer seg om bare ett flertydig ord, ville man i beste fall bare kunne generalisere til den forventede oppførsel av dette ordet i et

større korpus (og ikke til resultatene ved andre ord). Et siste viktig moment er at de tre klassifikatorene ikke bare trenes på korpora av ulike størrelser, men i tillegg er også den kvantitative sikkerheten i presisjonsresultater ulik mellom de tre korpusene: TiMBL kalkulerer en klassifikators presisjon basert på det anvendte korpusets "fasit". Dette betyr at siden K1 og (i noe mindre grad) K2 inneholder feilkilder i betydningstaggingen av korpuset, så kan det være at klassifikatoren egentlig ikke presterte nøyaktig i tråd med "fasiten" som dikteres av det på forhånd betydningstaggede korpuset.

Det siktes derfor heller mot en vurderende diskusjon av klassifikatorenes resultater så vel som å se *bak* de kvantitative resultatene som produseres av TiMBL. Det vil bli henvist til de kvantitative resultatene som suppleres av klassifikatorenes testresultater, men disse vil hovedsakelig tjene som grunnlag for å observere hovedtendenser i svarene på følgende spørsmål: Hvordan presterer hver av de tre klassifikatorene med henblikk på speilmetodens definerte betydninger av målordet? Vi forventer at siden betydningen CLAIM har flest treningseksempler, burde denne betydningen oftere kunne klassifiseres riktig enn de tre mindre frekvente betydningene; spesielt de to lavest frekvente "matrett"-betydningene COURSE og DISH. Imidlertid, dersom klassifikatoren ikke evner å "lære" å gjenkjenne noen av betydningene som distinkt fra de andre, må det vurderes hvorvidt dette kun er relatert til bare fattige data eller om kanskje speilmetodens betydningsskiller ikke er tilstrekkelig rimelige til å være anvendbar for trening av en klassifikator.

Vi vil også vurdere forskjeller mellom de tre klassifikatorene K1, K2 og K3: K1 inneholder "støy" (feilkilder) relatert til metodiske feil så vel som feilkilder som en følge av de praktiske, implementertingsmessige omgivelsene. Det er derfor naturlig å forvente at klassifikatoren K1 oftere foretar feil valg enn klassifikatorene K2 og K3. Dette leder oss til spørsmålet om hva resultatene antyder om oppgavens presenterte metode for ABT: Om vi anser K3s resultater som en "gullstandard" for hvor langt det er mulig å komme i treningen av en klassifikator med målordet *rett*Ns betydningskiller, er der da slik at klassifikatoren K2s metodiske feil spiller en stor rolle for resultatene?

Vi begynner i (4.4.2) med en analyse av resultatene til klassifikatoren K1. I (4.4.3) trenes en klassifikator K2 som skal antyde hvordan ABT-metodens feil står i forhold til de praktiske feil som følger av implementeringsmessige problemer i K1. I (4.4.4) følger treningen av en "gullstandard" klassifikator K3, som tjener som referansepunkt for en analyse av ABT-metoden som alternativ til manuell betydningstagging av et korpus.

4.4.2 Klassifikatoren K1. Vurdering av ABT-metodens treningsmateriale for klassifisering

Klassifikatoren K1 tilsvarer Lokal&Nøkkel17-klassifikatoren som ble presentert i (4.3.4), hvor det kun ble rapportert fra delresultatenes presisjon (antall korrekte klassifiseringer) og klassifikatorens prosentvise gjennomsnitt. Siden K1s prestasjonsresultater er kalkulert på basis av ABT-materialets oppførte betydningstag, gjør den relativt store mengden "støy" i K1s treningsdata at disse resultatene ligger svært lite til rette for en kvantitativ evaluering. Vi vil likevel ta utgangspunkt i TiMBLs kvantifiserende output, og se bak resultatene.

TiMBL supplerer en nyttig såkalt *confusion matrix* for hvert delresultat, som viser hvordan testmaterialets instanser ble klassifisert av maskinlæringsalgoritmen. Siden vi anvendte tidelt kryssvalidering som testmetode, får man ut ti slike matriser; vedlagt i appendiks (3). For illustrasjon gjengir (9) under matrisen til klassifikatoren K1s partisjon 1. I matrisens vertikale linje listes betydningstaggene som forekom i denne partisjonens testmateriale ifølge ABT-metodens "fasit": Tjue instanser betydningstagget som CLAIM, to instanser tagget som COURT og én instans tagget som DISH. Den horisontale linjen viser hvordan K1 valgte å klassifisere disse testinstansene. Vi ser at én forekomst av CLAIM ble galt klassifisert som COURT, mens de øvrige CLAIM-instansene ble korrekt identifisert. De to forekomstene av COURT ble også klassifisert riktig, mens DISH-instansen ble forvekslet med CLAIM.

(9) ***TiMBLs genererte confusion matrix* for klassifikatoren K1s testsett (partisjon) 1:**

partisjon 1	Antall korrekte: 21/23 Prosentvis presisjon: 91,30 %			
	claim	**court**	**course**	**dish**
claim	19	1	0	0
court	0	2	0	0
course	0	0	0	0
dish	1	0	0	0

I det følgende presenteres en vurdering av hvordan klassifikatoren K1 presterer med hensyn til hver av speilmetodens definerte betydningskategorier. For oversiktens del tillater vi oss her en felles oppsummering av hvordan klassifikatoren K1 presterte for hver av de fire betydningene, ved å slå sammen resultatene fra de ti delresultatenes matriser fra appendiks (3).[24] Tabell 4.4 oppsummerer hvor mange instanser der var av hver betydningskategori i følge ABT-korpuset, og hvor mange av disse som ble korrekt tagget ifølge ABT-materialets "fasit", det vil si betydningstaggene supplert av ABT-metoden. I siste kolonne er hver betydnings presisjon kalkulert relativt til betydningens totale antall instanser i ABT-materialet.

[24] Legg for ordens skyld merke til at korpusets 228 instanser ikke ble klassifisert på én gang slik det kan se ut av tabell 4: Testingen av K1 foregikk som vi husker via kryssvalidering. Med andre ord ble de 23 første instanser klassifisert i første test, de 23 neste i test nummer to osv., frem til de resterende 21 instanser i den siste testen.

Tabell 4.4: K1s klassifiseringsresultater sortert etter betydningskategori i ABT-materialet.

	Totalt antall	Riktig tagget*	Galt tagget*	Presisjon* %
claim	167	154	13	92,22
court	43	19	24	44,19
course	10	0	10	0,00
dish	8	5	3	62,50
TOTALT	228	178	50	78,07

* Ifølge betydningstaggene supplert av ABT-metoden

Vi ser at den store overvekten av treningsmaterialets eksemplifiserer betydningen CLAIM (167 av 228). Selv om ABT-metodens "fasit" ikke er helt presis, antyder tabell 4.4 at betydningen representert av CLAIM som oftest ble korrekt klassifisert i henhold til ABT-metodens betydningstagger. Tabell 4.4 viser at ifølge ABT-materialet ble 13 av CLAIM-instansene galt klassifisert.

For å illustrere problemer relatert til "støy" i ABT-materialet, skal vi se på to av tilfellene hvor en instans som ABT-metoden tagget som CLAIM ble klassifisert "galt" av K1. Den første av disse illustrerer hvordan "støy" i ABT-materialet kan påvirke klassifiseringen i gal retning. I (10) er testinstansen korrekt tagget i ABT-materialet, mens klassifikatoren K1 valgte en gal betydningskategori fordi den "nærmeste naboen" i treningsmaterialet var galt tagget av ABT-metoden. TiMBLs output for testinstansen inneholder følgende informasjon: Etter testinstansens 17 trekkverdier (kontekstord) følger ABT-metodens tilordnede betydningstag, og dernest klassifikatorens forslag. Krøllparentesen helt til slutt angir klassifikatorens vekting av hver betydningskategori som ble funnet som "nærmeste naboer" i treningsmaterialet. I (10) under kan vi altså lese at ifølge ABT-materialet skulle denne testinstansen tagges som CLAIM, mens TiMBLs klassifikator valgte COURSE. Ut fra krøllparentesen var dette den eneste "nærmeste nabo"-kategorien som ble funnet. TiMBL gir mulighet for innsyn i hvilke "nærmeste naboer" i treningsmaterialet som klassifikatoren har basert sitt valg av betydningskategori på. "Naboen" til (10) er gjengitt i (11).

(10) gir,arbeidsmiljølov,rett,til,1,arbeidsmiljølov,omsorgspermisjon,stønad,fostring,adopsjon, rettighet,fosterbarn,fosterforeldre,foreldreansvar,barnevernsnemnd,fosterhjem,varsel, claim,course { course 1.00000 }

(11) ha,bare,rett,til,fødselspenger,sykepenger,fødselspenger,deltid,opptjene, fødselspengeperiode,søke,gjenoppta,helseinstitusjon,foreldreansvar,beregning,beregne, hovedregel,{ course 1 }

Som vi ser av (11), har K1 helt riktig identifisert en "nærmeste nabo" fra treningsmaterialet som er semantisk forenlig med testinstansen, i kraft av at (10) og (11) har de felles trekkverdiene *til* (som umiddelbar lokal høyrekontekst) og nøkkelordet *foreldreansvar*. Likevel blir klassifiseringen feil som en følge av en feilkilde i ABT-materialet. Det skal imidlertid kanskje bemerkes at dersom det totale treningsmaterialet var langt større, så er det ikke utenkelig at en galt betydningstagget "nærmeste nabo", som (11), kunne blitt "nedstemt" under klassifiseringen av andre, riktig betydningstaggede instanser.

Instansen (12) under illustrerer på sin side et tilfelle hvor ABT-metoden hadde tagget testinstansen galt som COURSE, i stedet for riktig CLAIM. ((12) eksemplifiserer en instans som egentlig ikke skulle kommet med i ABT-materialet, fordi *rett*N selv ikke korresponderte med et ord som tilhørte en av *rett*Ns betydningspartisjoner.) Selv om klassifikatoren faktisk

korrekt har identifisert testinstansen som CLAIM, blir instansen altså likevel regnet som å ha "feil" betydningskategori ifølge ABT-materialets "fasit".

(12) ha,bare,rett,til,fødselspenger,sykepenger,fødselspenger,deltid,opptjene, fødselspengeperiode,søke,gjenoppta,helseinstitusjon,foreldreansvar,beregning, beregne,hovedregel,course,claim {claim 1.00000}

Klassifikatoren K1s valg av betydningstaggen CLAIM var basert på følgende, forenlige "nærmeste nabo" i treningsmaterialet:

(13) ha,samme,rett,til,permisjon,yrkesaktiv,fødselspenger,rette,omsorgspermisjon,forslag, forelder,vikar,tidsbegrense,arbeidsavtale,permisjonstid,strekk,unntak,{ claim 1 }

Disse eksemplene illustrerer altså hvordan "støy" i ABT-materialet kan påvirke klassifiseringen. Imidlertid er der også tilfeller hvor feilen ikke bunner i "støy", men i at kontekstinformasjonen mellom testinstansen og tilgjengelige treningsinstanser rett og slett ikke var tilstrekkelig, noe som er naturlig å relatere til fattige treningsdata.

Tabell 4.4 ovenfor antyder videre at den ABT-materialets nest mest frekvente betydningen, COURT, oftere tagges galt enn riktig ifølge betydningstaggene som ABT-metoden oppgir. Ifølge K1s ti matriser i appendiks (3) er det slik at av de 24 gangene en COURT-instans var klassifisert galt, ble det 23 ganger foreslått CLAIM, og bare én gang en "matrett"-betydning, nemlig COURSE. En kontroll av instansen som ble klassifisert som COURSE framfor COURT avslørte imidlertid at instansen var et eksempel på "støy" i ABT-materialet. Som vi ser i (14) under, ble instansen klassifisert som COURSE funnet å ligne på en instans (15) som manuelt ville blitt assosiert med betydningen CLAIM. Imidlertid var (15) en av instansene hvor en tilfeldig betydningstag ble generert pga. manglende ordparallellstilling, mens *rett*N selv ikke hadde en korrespondent som matchet *rett*Ns betydningspartisjoner. Fordi ordet *naturligvis* (det første trekket i (15)) var oversatt til engelsk *of course*, ble betydningstaggen COURSE altså generert på feil grunnlag. Når vi ser bak K1s kvantitative resultater, ser vi altså at når instanser av COURT er galt klassifisert, later CLAIM til å være et konsekvent valg.

(14) frifinnelse,kunne,rett,også,avsi,frifinnelse,B,nokon,blodprøve,gransking, hovedforhandling,avsi,samtykke,farskap,også,ordlyd,paragraf,court,course { course 1.00000 }

(15) naturligvis,hatt,rett,også,-,intelligens,egenhet,overaktiv,Sebastian,vanskelighet, legemliggjørelse,Såsnar,anfall,virkelyst,tilbaketrekking,veksle,tyrannisere,{ course 1 }

Strengt tatt er det imidlertid vanskelig å fastslå sikkert hvorfor klassifikatoren K1 ikke har forvekslet COURT med "matrett"-betydninger like ofte som med den andre "rettslige" betydningen, siden de to "matrett"-betydningene er svært lavfrekvente i materialet. Det er teoretisk mulig at dersom "matrett"-betydningene var sterkere representert i treningsmaterialet, så ville de også oftere vært representert som "nærmeste naboer" for testinstanser. Det framgår i den forbindelse av matrisene i appendiks (3) at når "matrett"-betydningene klassifiseres galt, så er de ikke forvekslet med hverandre men med de "rettslige" betydningene. Vi skal nå se nærmere på de to "matrett"-betydningene.

Ved en manuell kontroll viste de to "matrett"-betydningene seg faktisk å være mindre frekvente i korpuset enn ABT-materialet antyder: Ifølge tabell 4.4 er der henholdsvis åtte og ti instanser av DISH og COURSE i ABT-korpuset. Basert på en manuell kontroll kom det frem at bare seks av de åtte DISH-instansene faktisk eksemplifiserer denne betydningen, mens de

øvrige to egentlig var eksempler på *rett*A som kom med i materialet ved en tilfeldighet. Likeledes var der bare tre instanser i hele treningskorpuset som skulle vært tagget som COURSE, mens de øvrige sju er galt taggede substantiver og feilaktig inkluderte forekomster av adjektivet *rett*N. Med andre ord er de to "matrett"-betydningene i virkeligheten bare representert av til sammen ni instanser i hele ENPC-korpuset. Når man tar dette minimale antallet eksempelinstanser i betraktning, og at ingen av betydningene heller ikke er feilfritt tagget i ABT-materialet, er det nærmest imponerende at fem av de seks korrekte ABT-instanser av DISH-betydningen ble korrekt identifisert av K1. Som eksempel gjengir (16) en av de korrekt klassifiserte instansene.

(16) og,en,rett,som,bare,trøffel,kalv,lapskaus,urt,lammestek,more,fantaisie,chef,gamling, bestilling,nikke,alltid,dish,dish { dish 1.00000 }

Som vi ser av informasjonen om funnede kategorier blant de "nærmeste naboer", ble instansen funnet å ha kun DISH-kategorien som "nærmeste nabo" i treningsmaterialet. Den "nærmeste naboen" er gjengitt i (17) under. Som vi ser, har testinstansen og treningsinstansen to trekk felles, nemlig det første ordet umiddelbart til venstre og høyre for målordet (*..en rett som..*).

(17) -,en,rett,som,opprinnelig,kapama,arni,bouillabaisse,fortreffelig,i_går,roe,Kalamata, Peloponnes,mager,lammekjøtt,steke,hvitvin,{ dish 1 }

Når det gjelder COURSE-betydningen, ble ingen av de tre faktiske instansene av betydningen gjenkjent av K1. Dette kan ikke sies å være overraskende, gitt betydningens høyst sparsomme treningsgrunnlag. Alle ABT-metodens ti oppgitte instanser av COURSE ble klassifisert som enten CLAIM eller COURT av K1. Den ble altså aldri forvekslet med den andre "matrett"-betydningen i det endelige valget av betydningskategori, noe som først og fremst later til å bunne i de svært fattige data som suppleres både for begge "matrett"-instansene. Instansen (18) under illustrerer et tilfelle av COURSE-betydningen hvor alle fire betydningskategorier ble vurdert. K1 fant til sammen 13 "nærmeste naboer" i treningsmaterialet: 9 som sammenfalt med CLAIM , 2 med COURT, og bare 1 som lignet på henholdsvis COURSE og på DISH.

(18) komme,neste,rett,-,-,halvmeterlang,brødskive,tallerken,ansjos,måltid,uforglemmelig, postei,hare,villsvin,trost,tykkfallen,skinketerrin,course,claim { claim 9.00000, court 2.00000, course 1.00000, dish 1.00000 }

En nærmere analyse av de tretten "naboene" viste at de ble funnet å ligne testinstansen i (18) fordi målordet i alle instanser har trekkene ”-,-” som lokal høyrekontekst. (Med andre ord var målordet siste ord i den aktuelle setningen.) Dette antyder for det første nok en gang at den lokale konteksten later til å spille størst rolle. Men i tillegg til dette, ser vi for det andre at klassifikatoren altså ikke var sikker på hvordan (18) skulle kategoriseres, siden alle fire betydninger ble vurdert. Ettersom betydningen CLAIM er sterkt overrepresentert i treningsmaterialet i forhold til de øvrige tre betydningene (og spesielt i forhold de to "matrett"-betydningene), er det ikke usannsynlig at CLAIMs langt større totale mengde av eksempelkontekster vil gjøre at denne betydningen generelt favoriseres.

Som en oppsummering avtegner følgende bilde seg av klassifikatoren K1, som er trent på det automatisk betydningstaggede materialet. K1s største problem later til å være mangelen på et større treningsmateriale for å avgjøre hvilken betydning en instans av målordet skal tilordnes. Dette gjenspeiles først og fremst i at den mest frekvente betydningen CLAIM blir klassifisert med større prosentvis presisjon enn de øvrige tre betydningene (Tabell 4.4). Der er

også en tendens til at de tre mindre frekvente betydningene forveksles med CLAIM når de er galt klassifisert. Vi har videre sett at den relativt store andelen "støy" i ABT-materialet også skaper problemer for klassifikatoren. Det primære problemet med "støy" er at den kan påvirke klassifiseringen av andre instanser, slik vi for eksempel så i (10-11) og (14-15). Hvis en korrekt identifisert treningsinstans er assosiert med en tilfeldig (og gal) betydningstag, bringes denne feilen i ABT-materialet altså videre i klassifiseringen av nye instanser. Det skal likevel påpekes at hver betydnings presisjon i tabell 4.4, med unntak av course, er bedre enn tilfeldig. (Et tilfeldig valg av betydningskategori gir en 25 % sjanse for hver av kategoriene).

Vi skal nå gå over til eksperimentet med å trene en klassifikator K2, hvor bare "støy" som er en konsekvens av metoden for ABT beholdes mens feil av mer praktisk art lukes bort. K2 kan dermed antas å gi et sannere bilde av ABT-metodens potensial for å ekstrahere finitte betydningstaggede korpora til bruk i overvåket WSD.

4.4.3 Klassifikatoren K2. Rollen til metodiske feilkilder i ABT-materialet

I evalueringen av det automatisk betydningstaggede materialet i (3.4.1) ble det lagt frem tre typer feilkilder i ABT-metodens resulterende materiale:

I: Feilkilde i selve metoden for automatisk betydningstagging
II: Manglende ordparallellstilling
III: Svakheter ved preprosesseringen + manglende ordparallellstilling (uønskede instanser)

Med feilkilder i selve ABT-metoden mener vi feil hvor det implementerte systemet for ABT-metoden helt riktig har identifisert en match mellom en betydningspartisjon og *rett*Ns korrespondent og tilordnet instansen en betydningstag, men hvor resultatet likevel synes intuitivt galt. (For eksempel tilfellet hvor en "matrett"-betydning metodisk sett korrekt tildeles betydningstaggen CLAIM fordi oversettelseskorrespondenten *order* tilhører CLAIMs betydningspartisjon.) Siden den presenterte ABT-metoden prinsipielt forutsetter tilgang på et flerspråklig korpus som er parallellstilt på ordnivå, regner vi det som egen type feil (II) når betydningstaggingen blir feil som en følge av at ENPC bare er parallellstilt på setningsnivå. Denne typen feil oppstår når en betydningstag genereres på basis av andre ord enn selve målordets korrespondent. Den tredje typen feil berører hva vi kalte *uønskede instanser*, det vil si forekomster av andre lemmaer enn *rett*N. Det forekommer som vi husker at adjektivet *rett* feilaktig inkluderes i det betydningstaggede materialet, som en følge av svak preprosessering av ENPC kombinert med mangelen på ordparallellstilling.

Den manuelle evalueringen av ABT-materialet i (3.4.1) påviste først og fremst feil av type II og III, altså feil som ikke har en direkte tilknytning til selve ABT-metoden, og vurderingen av K1 viste at disse feilene av og til påvirker klassifiseringen i gal retning. For et bedre bilde av selve metodens anvendelighet, vil vi nå tenke oss de "ideelle" implementeringsmessige omgivelser at vi har et korrekt (manuelt korrigert) preprosessert parallellkorpus, som er lenket på ordnivå (framfor på setningsnivå, som ENPC).

Korrigeringen av ikke-metodiske feil ble utført etter følgende kriterier. Vi tar utgangspunkt i ABT-materialet, slik det har blitt implementert og presentert i denne oppgaven. Om vi antar at parallellkorpuset var korrekt preprosessert (lemmatisert og pos-tagget), ville feil av type III lukes ut av oppgavens presenterte metode. Altså luker vi ut de 22 forekomstene av adjektiv-*rett* som feilaktig ble inkludert i materialet. Feilene av type II omfatter instanser av det riktige lemmaet, *rett*N. Om vi antar at vi hadde tilgang på en adekvat ordparallellstilling, ville feil av type II lukes ut av systemet fordi disse instansene av målordet

ikke ble tilordnet en betydningstag ut fra sin oversettelseskorrespondent. Derfor luker vi også ut de instanser av *rett*N som ble tagget på grunnlag av et tilfeldig lemma i det motstående språket, som ikke korresponderte med *rett*N. Vi står da igjen med et treningskorpus hvor vi lar ABT-metodens metodisk relaterte feil stå som de er, mens feil av type II og III er luket bort. Mens det opprinnelige ABT-korpuset besto av 228 instanser av *rett*N, inneholder det delvis korrigerte treningsmaterialet nå 198 treningsinstanser.

Det ble så trent en klassifikator K2 på dette delvis korrigerte treningsmaterialet, på samme måte som treningen av K1 foregikk: Ved tidelt kryssvalidering med TiMBLs læringsalgoritme IB1, og med kontekstvinduet Lokal&Nøkkel17. Vi skal først for sammenligningens skyld angi K2s delresultater, stilt opp mot K1s resultater:

Tabell 4.5: Klassifikatoren K1 (trent på det komplette ABT-materialet) og K2 (trent på et ABT-korpus med kun metodisk relaterte feil)

Klassifikator	**K1**	**K2**
Delresultater (antall korrekte av totalt antall instanser i hvert testsett)	21/23	19/20
	22/23	20/20
	18/23	17/20
	14/23	15/20
	18/23	14/20
	8/23	11/20
	18/23	17/20
	20/23	17/20
	21/23	52/20
	18/21	18/18
Totalt antall korrekte	178/228	168/198
Gj.snittlig presisjon %	**78,10 %**	**84,85 %**

Med forbehold om unøyaktigheter i det kvantitative resultatet, later K2 til å prestere betraktelig mye bedre enn K1. Dette var også som forventet, siden sistnevnte klassifikator ikke inneholder feil av typene II og III i sitt treningsmateriale. For å få et bedre bilde av hva som ligger bak delresultatene, vil vi nå gå nærmere inn på hvilke hva som er klassifikatoren K2s problemområder i forhold til K1. Som i forrige delkapittel vil vi ikke her gå inn på hvert delresultats *confusion matrix*, disse er vedlagt i appendiks (4). Tabell 4.6 oppsummerer hvordan de 198 instansene som var del av K2s klassifiseringsoppgave fordelte seg mellom betydningskategoriene, og angir prosentvis presisjon for hver betydningskategori. (Men legg merke til at dette utgjør summen av ti separate delresultater uten å spesifisere hvordan betydningene fordelte seg mellom de ti delresultatene, og de ti adskilte matrisene i appendiks (4) er således mer presise.)

Tabell 4.6: K2s klassifiseringsresultater sortert etter betydningskategori i ABT-materialet korrigert for ikke-metodiske feil.

	totalt	Riktig tagget*	Galt tagget*	Presisjon* %
claim	148	138	10	93,24
court	41	25	16	60,98
course	3	0	3	0,00
dish	6	5	1	83,33
TOTALT	198	168	30	84,85

* ifølge det delvis korrigerte ABT-materialets data om hva som er ønsket betydningstag

Tabell 4.6 viser at det totale antallet instanser for hver betydning nå er endret i forhold til K1s treningsmateriale: 19 instanser tagget som CLAIM er fjernet, likeledes 2 instanser tagget som COURT, og antallet instanser av de to "matrett"-betydningene er nå i samsvar med antallet instanser som i ABT-materialet var korrekt betydningstagget med en av "matrett"-betydningene (tre og seks for henholdsvis COURSE og DISH). For COURT og CLAIM er der imidlertid fremdeles åtte tilfeller av metodisk relaterte feil i materialet. I tillegg til instansen av "matrett"-betydningen som er galt tagget som CLAIM, er der noen instanser som vi intuitivt ville tilordnet betydningen COURT ved manuell tagging av korpuset. Eksempel fra treningsmaterialet:

(19) i,norsk,rett,være,myndighetsalder,myndighetsalder,opphøre,foreldreansvar,Norge, skille,ung,lagtingslov,Bjarkøyretten,Magnus,Lagabøter,landslov,aldersgrense,claim

Intuitivt ville vi kanskje foretrekke å tagge instansen i (19) som COURT. Ettersom målordet i (19) imidlertid korresponderte med engelsk: *..under Norwegian **law**, only..*, og siden *law* er innordnet betydningspartisjonen CLAIM, tagges instansen metodisk sett helt korrekt som CLAIM.

Tabell 4.6 ovenfor viser at i henhold til ABT-metodens genererte betydningstagger, har ti instanser av betydningen CLAIM blitt galt klassifisert av K2. Matrisene i appediks 4 viser at CLAIM så å si konsekvent forveksles med COURT. Som (20) under eksemplifiserer, bunner de fleste av disse instansene i "feil" i ABT-metoden. Instansen (20) er et av tilfellene hvor ABT-metoden metodisk riktig genererer CLAIM som betydningstag, fordi målordet korresponderte med engelsk *law*. Siden setningen dreier seg om "i gammel norsk rett" med referanse til rettssystem, ville imidlertid kanskje COURT intuitivt være mer ønskelig. Derfor er det vanskelig å si at klassifikatoren gjør direkte "feil" i å velge COURT framfor CLAIM. På den andre siden skal det også bemerkes at det selv manuelt er vanskelig å avgjøre om ABT-metodens valg av betydningen CLAIM egentlig er så mye mer feil enn COURT. K2s to funnede "nærmeste naboer" er for ordens skyld gjengitt i (21).

(20) gammel,NIL,rett,skulle,bare,skille,voksen,pubertet,grensetrekning,tidlig,relasjon, våpenfør,myndig,den_gang,ungdomstid,barn,Norge,claim,court { court 2.00000 }

(21) -,-,rett,skulle,også,B,farskapsforelegg,påtegning,farskap,samtykke,mor,oppgitt, tilstrekkelig,fastslå,pkt.,frifinnelse,anse,{ court 1 }

være,at,rett,skulle,avgjøre,farskapssak,regel,pålegge,B,innhente,omkostning,avgi,tie, tidlig,påberope,Rl,aktelse,{ court 1 }

Vi observerer altså en temmelig klar tendens til at K2, som K1, klassifiserer CLAIM-instanser relativt sikkert. Som det ble påpekt også for K1, bunner dette sannsynligvis i at denne betydningen har de langt fleste eksempelinstansene i korpuset. Vi observerer videre at når K2 klassifiserer en CLAIM-instans av målordet "galt" ifølge ABT-metodens dikterte betydningstag, så er det betydningen COURT som K2 velger. Ettersom (20) ovenfor illustrerer at det ikke alltid er intuitivt enkelt å skille mellom COURT og CLAIM, må dette sies å være "feil" som ikke synes svært graverende.

K1s tendens til å forveksle COURT med CLAIM (ifølge ABT-materialet) gjentar seg i K2s klassifisering. Mens K1 samlet sett klassifiserte en overvekt av COURT-instansene galt, klassifiserer imidlertid K2 en majoritet av instansene korrekt. Det skal her bemerkes at både K1 og K2s partisjoner med fordel kunne vært randomisert: Ved begge klassifikatorer tar den sjette testpartisjonen bort ca. halvparten av det totale antallet treningsinstansene som korpuset supplerer for COURT. K2 klassifiserer likevel en knapp majoritet av COURT-instansene

riktig også i dette sjette testsettet (mens K1 bare klassifiserte 6 av 21 COURT-instanser riktig i dette settet, altså et klart mindretall).

I K2s treningsmateriale er "matrett"-betydningene betydningstagget med 100 % presisjon, det vil si at alle instanser tagget som COURSE eller DISH er faktiske eksempler på "matrett"-betydningen. I tillegg er der også, som vi husker, en instans som *skulle* vært tagget med en "matrett"-betydning, men som i stedet metodisk korrekt ble tagget som CLAIM (fordi *rett*N korresponderte med engelsk *order* i "matrett"-betydningen). K1 gjenkjente ikke denne sistnevnte instansen som en "matrett"-betydning, og foreslo å klassifisere den som CLAIM. For K1 var dette kanskje heller ikke overraskende, siden de to "mat"-betydningene hadde store feilkilder i sitt allerede sparsomme treningsmateriale. Det viste seg at heller ikke K2s treningsmateriale ga grunnlag for å klassifisere denne instansen korrekt, og både K1 og K2 har klassifisert denne instansen på identisk grunnlag. Som vi ser av (22) og K2s funnede "nærmeste naboer" i (23) under, ble der riktignok funnet én "nærmeste nabo" assosiert med "matrett"-betydningen COURSE. Imidlertid ble denne nedstemt av at der var *to* nærmeste naboer tilknyttet CLAIM-betydningen, som dermed ble valgt foran både COURSE og COURT-betydningen som også ble vurdert.

(22) når,en,rett,være,klar,verdig,forsømme,forsinkelse,likevel,makte,servitør,madame,klokke, øyebryn,traske,mumle,j'arrive,claim,claim { claim 2.00000, court 1.00000, course 1.00000 }

(23) en,naturlig,rett,være,han,avkall,løgn,forråde,manndom,gjennomskue,illusjon,forsyne, oppvigling,streik,cricket-resultat,ukeblad,desinformasjon,{ claim 1 }

-,Senjen,rett,være,bare,nytte,dømme,Horesønner,rope,Ila,gylden,mellomstasjon, ubetydelig,lutrygga,strikkejakke,blodsmak,dødelig,{ court 1 }

omløpstid,-,rett,være,basere,omløpstid,innrette,sted,Ami,Bel,effektuere,prefabrikat, dypfrossen,delikatesse,pommes_frites,ingrediens,frityrolje,{ course 1 }

i,norsk,rett,være,myndighetsalder,myndighetsalder,opphøre,foreldreansvar,Norge,skille, ung,lagtingslov,Bjarkøyretten,Magnus,Lagabøter,landslov,aldersgrense,{ claim 1 }

Det er et empirisk spørsmål om flere treningseksempler for "matrett"-betydningene kunne bidratt til at denne instansen ble riktig klassifisert. Siden hovedpoenget i denne diskusjonen imidlertid er å anskueliggjøre eventuelle svakheter ved å anvende oversettelseskorrespondanser som grunnlag for betydningstagging av et korpus til bruk i en klassifiseringsoppgave, kan vi konkret formulere spørsmålet som følger: Gjør en galt betydningstagget instans i treningsmaterialet, som (22), at *andre* "matrett"-relaterte instanser av målordet blir feil? Dette vil tas opp i den nå følgende evalueringen av K2s klassifisering av de to "matrett"-betydningene.

Av materialets seks instanser av DISH, ble fem korrekt klassifisert. Den siste ble galt klassifisert som CLAIM som en følge av fattige treningsdata. Betydningskategorien COURSE har kun tre eksempelinstanser i korpus, og K2 lykkes ikke i å klassifisere noen av dem: To av dem klassifiseres som CLAIM, mens den siste foreslås å være en instans av COURT. Grunnen til dette later kort sagt til å være fattige treningsdata. I ett av tilfellene, gjengitt i (24), fant K2 riktignok én "nabo" som vi gjenkjenner som "matrelatert", nemlig den siste "naboen" i (25). Dette er instansen som korresponderte med *order* og som følgelig ble tagget som CLAIM av ABT-metoden. Vi ser imidlertid også at selv hvis denne "naboen" hadde vært korrekt tagget med en "matrett"-betydning, ville likevel betydningskategorien CLAIM vært i overtall (med to "naboer"). Følgelig ville (24) likevel blitt klassifisert som CLAIM. Slik sett kan vi si at det er mengden treningsmateriale snarere enn en svakhet ved ABT-metodens betydningstagger som gjør utslaget her.

(24) omløpstid,-,rett,være,basere,omløpstid,innrette,sted,Ami,Bel,effektuere,prefabrikat,
dypfrossen,delikatesse,pommes_frites,ingrediens,frityrolje,course,claim
{ claim 3.00000, court 1.00000 }

(25) en,naturlig,rett,være,han,avkall,løgn,forråde,manndom,gjennomskue,illusjon,forsyne,
oppvigling,streik,cricket-resultat,ukeblad,desinformasjon,{ claim 1 }

-,Senjen,rett,være,bare,nytte,dømme,Horesønner,rope,Ila,gylden,mellomstasjon,
ubetydelig,lutrygga,strikkejakke,blodsmak,dødelig,{ court 1 }

i,norsk,rett,være,myndighetsalder,myndighetsalder,opphøre,foreldreansvar,Norge,skille,
ung,lagtingslov,Bjarkøyretten,Magnus,Lagabøter,landslov,aldersgrense,{ claim 1 }

når,en,rett,være,klar,verdig,forsømme,forsinkelse,likevel,makte,servitør,madame,klokke,
øyebryn,traske,mumle,j'arrive,{ claim 1 }

Som en oppsummering har vi sett at klassifikatoren K2 presterer overraskende mye bedre enn K1: Med forbehold om unøyaktigheter i beregningen av antall riktig og gale klassifiseringer, har K2 sammenlagt en gjennomsnittlig presisjon på 84,85 %, mot K1s 78,10 %. Det ble estimert i den manuelt baserte evalueringen av ABT-materialet i (3.4.1) at med alle tre typer feil har treningsmaterialet en presisjon 83,3%. Til sammenligning er K2s treningskorpus manuelt korrigert for å reflektere hvordan selve ABT-metoden ville yte under ideelle implementeringsmessige omstendigheter (med tilgang på et tilfredsstillende preprosessert og ordparallellstilt korpus). Selve treningskorpuset inneholder da åtte metodisk relaterte feil, hvilket tilsvarer at ABT-metoden tagger instanser med en presisjon på 96 %. Det er derfor oppløftende å se at K2s forbedrede treningsmateriale mot K1s også later til å gjenspeiles i resultatene.

K2s materiale er, som K1s, temmelig skjevt fordelt mellom de fire betydningskategoriene speilmetoden utledet for *rett*N. Det var derfor kanskje som forventet at det var den mest frekvente betydningen CLAIM som også oftest ble riktig klassifisert. Det er vanskelig å trekke "sikre" konklusjoner ut fra et treningsmateriale som for det første totalt sett er begrenset av en sparsom treningsmengde, og hvor de ulike betydningskategoriene for det andre er såpass skjevt fordelt. I tillegg til disse to faktorene har vi også sett at resultatene kan påvirkes av at testingen av klassifikatoren har blitt utført ved tidelt kryssvalidering. Den observerbare tendensen synes like fullt å være at når klassifikatoren tar feil av de to "rettslig" relaterte betydningene, så forveksles disse med hverandre og svært sjelden med "matrett"-betydningene. Omvendt ser vi at de to "matrett"-betydningene derimot sjelden forveksles med hverandre, snarere blandes de i stedet sammen med de "rettslige" betydningene" (spesielt den totalt mest frekvente CLAIM). Det er imidlertid ikke utenkelig at de to mest frekvente betydningskategoriene har en generell tendens til å favoriseres, siden de er overrepresentert i materialet i forhold til "matrett"-betydningene. Hvis mengden av eksempler på hver betydning var jevnbyrdige, er det således godt mulig at et helt annet mønster ville avtegne seg.

Uten at hver eneste klassifiserte instans er gjennomgått, har det ikke blitt identifisert noen eksempler hvor K2 klassifiserer en testinstans galt som en følge av metodisk galt betydningstaggede treningsinstanser. Derimot så vi at K1, som trenes på et materiale med alle tre definerte typer feil, av og til foretar feil valg på grunn av feil i treningsmaterialet. Siden der ikke har blitt funnet tilsvarende feil i de gjennomgåtte resultatene for K2s klassifisering, synes det rimelig å anta at de metodiske feilene faktisk er såpass få at de ikke påvirker K2s resultater i særlig grad.

Vi vil nå gå videre til å evaluere klassifikatoren K2 mot en klassifikator K3 som er trent på manuelt korrigerte instanser av alle forekomster som finnes av *rett*N i ENPC.

4.4.4 Klassifikatoren K3. Manuelt betydningstagget treningskorpus

I dette delkapittelet vil vi trene en klassifikator K3 som er ment å tjene som referansepunkt for hvor godt det ville være mulig for en klassifikator å prestere, gitt de instanser av målordet som er tilgjengelig i ENPC. For å undersøke dette har alle instanser av *rett*N i ENPC blitt tagget manuelt med speilmetodens utledede betydningsskiller. Dette korpuset består da av alle de instanser som ABT-metoden forkastet, samt en ny versjon av det opprinnelige ABT-materialet. Denne versjonen er korrigert for feil på følgende måte: Alle uønskede instanser av *rett*A (feil av type III) er luket ut. Videre er alle instanser av det faktiske målordet, med feil av type II (manglende ordparallellstilling) og I (metodiske feil), manuelt korrigert. Dette gir til sammen et treningskorpus på 398 instanser. Som klassifikatorene K1 og K2, blir K3 trent ved tidelt kryssvalidering. Antall korrekt klassifiserte instanser i hvert av de ti testsettene er gjengitt i tabell 4.7 under.

Tabell 4.7: Resultater for klassifikatoren K3, trent ved tidelt kryssvalidering på et manuelt tagget korpus

Klassifikator	**K3**
Delresultater (antall korrekte av totalt antall instanser i hvert testsett)	37/40
	37/40
	17/40
	33/40
	40/40
	37/40
	38/40
	36/40
	38/40
	35/36
Totalt antall korrekte	348/396
Gj.snittlig presisjon %	**87,88**

Siden størrelsen på K2 og K3s treningskorpora (henholdsvis 198 og 398 instanser av målordet) er enda mer ulik enn forskjellen mellom K1 og K2, kan K2 vanskelig sammenlignes direkte mot K3 basert på kvantitative mål. Det minnes likevel om at klassifikatoren K2 ble funnet å klassifisere til sammen 168 av sine 198 instanser av målordet korrekt, hvilket prosentvis tilsvarer en presisjon på 84,85 %. Som vi ser, klassifiserer K3 til sammen 348 av 396 instanser korrekt, hvilket gir en sammenlagt presisjon på 87,88 %. Selv om vi skal huske at K2 og K3s resultater ikke er direkte sammenlignbare, er det nesten overraskende at K3 prosentvis ikke presterer enda bedre i forhold til K2 enn hva tabell 4.7 viser. K3 er som vi husker ment å reflektere hvor godt det er mulig for en klassifikator å prestere, gitt en manuelt verifisert betydningstagging av alle ENPCs instanser av *rett*N. Med K3s resultater som referansepunkt synes faktisk K2s resultater med et automatisk betydningstagget materiale ganske lovende.

Hvert testsetts *confusion matrix* er vedlagt i appendiks (5). K3s treningsmateriale er sammensatt slik at de 198 instansene som også ble klassifisert av K2 kommer først. Med

andre ord består K3s partisjoner nummer 1 til 5 av de samme instansene som K2, minus de to siste instansene i K3s partisjon nummer 5. (Disse to siste instansene er henholdsvis korrekt klassifiserte instanser av COURT og CLAIM.) Som en oppsummering av matrisene i appendiks (5), gir tabell 4.8 under et overblikk over hvordan de totalt 396 instansene som inngikk i K3s klassifiseringsoppgave fordelte seg mellom betydningskategoriene, og angir prosentvis presisjon for hver betydningskategori.

Tabell 4.8: K3s klassifiseringsresultater sortert etter betydningskategori i det manuelt taggede materialet

	totalt	Riktig tagget	Galt tagget	Presisjon %
claim	333	323	10	97,00
court	52	23	29	44,23
course	4	0	4	0,00
dish	7	2	5	28,57
TOTALT	396	348	48	87,88

Tabell 4.8 illustrerer tydelig den skjeve fordelingen mellom speilmetodens fire betydningskategorier for *rett*N i ENPC. Tabellen kan kanskje også sies å antyde at mengden av treningsdata for hver kategori gjenspeiles i K3s presisjon innenfor hver betydningskategori. I det følgende presenteres en kort analyse av K3s resultater for hver kategori.

Kategorien CLAIM har den definitivt største treningsmengden: Om vi regner i relativ frekvens mellom de fire betydningskategoriene utgjør instansene av CLAIM hele 84,1 % av det totale antallet på 396 instanser i korpuset. Det er således ikke overraskende at denne kategorien også blir oftest korrekt klassifisert. (Til sammen for alle ti delresultater klassifiseres CLAIM med en presisjon på 97 %.) En nærmere analyse av feilene innenfor denne kategorien viser det samme mønsteret som ble observert for K2. Når CLAIM klassifiseres galt, velges nesten konsekvent den andre "rettslige" betydningen, COURT. I ett tilfelle valgte klassifikatoren K3 i stedet DISH. Denne instansen ble også galt klassifisert av både K1 og K2.

Kategorien COURT representerer 52 av de totale målordsinstansene i korpuset (13,1 % av materialet, målt i relativ frekvens). Til forskjell fra K2 ble denne betydningen oftere klassifisert galt enn riktig av K3. Prosentvis klassifiseres betydningen med en presisjon på 44,23 %. K3s sammenlagte resultater for COURT er altså svakere enn K2s, hvor COURT til sammen ble korrekt klassifisert i 60,98 % av tilfellene. Som ved K2 later dette imidlertid til en viss grad til å komme av at tilfeldigheter i partisjoneringen av K3s testsett. Den tredje partisjonen av K3s materiale (appendiks (5)) viser at dette testsettet tilfeldigvis inneholdt hele 27 av COURTs totalt 52 eksempelinstanser. Følgelig inneholder treningsmaterialet for dette tredje testsettet under halvparten av alle instansene som eksemplifiserer betydningen. Matrisen viser at i dette testsettet ble 20 av de 27 instansene galt klassifisert som CLAIM. Når dette er sagt, ser vi imidlertid også av matrisene til partisjon 1, 8 og 9 at klassifikatoren også generelt synes å ha en liten tendens til å velge CLAIM framfor korrekt COURT. Gitt at der er så stor forskjell mellom mengden treningsmateriale for de to "rettslige" betydningene, er sannsynligvis den ujevne fordelingen i treningsmengde grunnen til at de to betydningene oftere forveksles i tilfellet COURT enn i tilfellet CLAIM. Det later likevel til å være en klar tendens til at når de to "rettslige" betydningene er galt klassifisert, så forveksles de med hverandre. Dette kan kanskje tolkes som å gjenspeile at forskjellen mellom dem ikke er like klart kontrastiv som forskjellen mellom betydningene "rettslig" og "mat" for målordet. Som påpekt er dette imidlertid vanskelig å avgjøre sikkert, siden "matrett"-betydningene er så lavt frekvente.

Som det fremgår av tabell 4.8, er det ikke bare i ABT-materialet at de to "matrett"-betydningene er sterkt underrepresentert i forhold til de "rettslige" betydningene. COURSE er instansiert fire ganger i hele ENPC, mens betydningen DISH eksemplifiseres av syv instanser. Det er derfor ikke overraskende at klassifikatoren K3, som K2, har problemer med å identifisere disse betydningene. Verken K2 eller K3 (og heller ikke K1) greide å klassifisere noen av instansene som var betydningstagget som COURSE, hvilket egentlig heller ikke kan forventes med så fattige data. Med hensyn til betydningen DISH, lyktes imidlertid K2 i å gjenkjenne fem av de seks instansene som K2s materiale ga tilgang på. K3 har tilgang på én instans mer, men lykkes bare i to av tilfellene å klassifisere betydningen riktig. Dette kan forklares ved to faktorer: For det første spiller det åpenbart en rolle at begge de to "matrett"-betydningene totalt sett er lite frekvente i K3s materiale (til sammen elleve instanser). For de andre viser matrisene i appendiks (5) at av disse elleve instansene forekom sju i det samme testsettet, partisjon 4 (fire DISH og tre COURSE). Følgelig har dette testsettet bare fire treningseksempler å basere klassifiseringen på (tre DISH og én COURSE). I dette testsettet lyktes K3 i klassifisere én instans av DISH korrekt, mens de øvrige seks "matrett"-betydningene klassifiseres som enten CLAIM eller COURT. En nærmere analyse viste imidlertid at klassifikatoren hadde identifisert "mat"-relaterte "nærmeste naboer" ved fem av de seks "mat"-instanser som ble galt klassifisert i dette testsettet. Imidlertid ble de funnede "matrett"-relaterte "naboer" hver gang nedstemt, ved at der var flere "nærmeste naboer" som eksemplifiserte enten CLAIM eller COURT (og hver gang spilte for øvrig også instansenes lokale høyrekontekst inn). Selv om klassifikatoren altså ikke lyktes i velge riktig betydningskategori, later den likevel til å ha vært inne på riktig klassifikator i de fleste tilfeller.

4.5 Oppsummering

Evalueringen av de tre klassifikatorene har på den ene siden vist at alle tre lider under at deres respektive treningsmateriale for det første er sparsomt totalt sett, og kanskje enda viktigere at de fire betydningskategoriene er svært ulikt fordelt med hensyn til relativ frekvens i materialet. Med tanke på ABT-metoden som et alternativ til manuell tagging, antyder den relativt lille forskjellen i prestasjon mellom K2 og K3 at dette problemet ikke er direkte forbundet med selve metoden for automatisk betydningstagging.

På den andre siden kan man innvende at spørsmålet om treningsmengde per betydning til en viss grad *er* å anse som et prinsipielt spørsmål relatert til denne oppgavens presenterte metode for automatisk ekstrahering av et betydningstagget treningskorpus: Ved manuell betydningstagging er det prinsipielt ikke noe i veien for å ekstrahere treningsinstanser fra flere ulike korpora, for å sørge for at hver betydning av målordet får like mange (eller i hvert fall over et definert minstemål) treningsinstanser å lære fra. For anvendelse i større skala (det vil si for flere ord enn denne oppgavens testlemma), forutsetter imidlertid oppgavens presenterte ABT-metode et parallellkorpus som er tagget for pos, lemmatisert og parallellstilt på ordnivå. Dermed kan vi si at metoden prinsipielt er nødt til å forholde seg kun til de treningseksempler per betydning som foreligger i det anvendte parallellkorpuset. (Eller teoretisk i flere parallellkorpora, så sant de tilfredsstiller metodens forutsetninger).

Slik sett kan det karakteriseres som et metodisk relevant problem at klassifikatorens resultater er synlig påvirket av variasjon i treningsmengde per betydning. K3 viste hvor langt det var mulig å komme i treningen av en klassifikator for *rett*N, basert på de instanser av målordet som foreligger i ENPC. Selv om det synes klart at den ujevne fordelingen mellom treningsmateriale påvirker klassifiseringen av de mindre frekvente betydningene, er det

likevel oppløftende å se at ABT-metodens klassifikator K2 ikke presterer markant svakere enn K3.

Eksperimentene antyder at selve metoden for automatisk betydningstagging har et godt potensial som alternativ til manuell betydningstagging av treningskorpora for overvåket WSD. Det største forskjellen mellom manuell og automatisk betydningstagging later til å være den totale mengden av treningsinstanser, siden ABT-metoden måtte forkaste knapt halvparten av ENPCs totale mengde instanser av *rett*N. Når det gjelder relativ fordeling mellom betydningskategoriene er imidlertid det manuelt og det automatisk taggede materialet så å si sammenfallende, og fra eksperimentene ser vi at klassifikatorene K2 og K3s heller ikke avviker stort fra hverandre resultatmessig.

Det ble videre observert at de metodiske feilene, til forskjell fra feil som en følge av de implementeringsmessige omgivelsene, ikke later til å påvirke klassifikatorens resultater i stor grad. Forskjellen i prestasjoner mellom K1 og K2 var nærmest overraskende stor, og antyder således at de feilkildene av praktisk art spilte en relativt stor rolle for K1s resultater. Ettersom K2 (som inneholdt metodisk relaterte feil i treningsmaterialet) ikke synes å ligge dårlig an i forhold til K3, later ikke denne typen feil til å ha spilt en like stor rolle for klassifiseringen.

Imidlertid må vi være klar over at testing på bare ett flertydig lemma ikke gir grunnlag for å trekke generaliseringene om ABT-metodens potensial for langt. Resultatene synes likevel å gi dekning for at der er god grunn til å teste ABT-metoden videre i større skala. Ettersom det arbeides med ordlenking av parallellkorpuset ENPC innenfor forskningsprosjektet "Fra Parallellkorpus til Ordnett", vil oversettelseskorrespondanser som input til speilmetoden kunne ekserperes automatisk. Forhåpentligvis gir dette i sin tur et mer omfangsrikt ekserperingsmateriale for å generere betydningspartisjoner, ikke minst for flere ord enn denne oppgavens ene testlemma. Siden problemet med manglende tilgang på betydningstaggede korpora for WSD er et stort problem innenfor WSD i dag, ville det være svært interessant å teste ut ABT-metoden med forbedrede korpusressurser.

5. Diskusjon og videre arbeid

I (1.4.4) ble det slått fast at en metode for automatisk betydningstagging som retter seg mot problemstilingen med mangelen på betydningstaggede korpora som bakenforliggende ressurs for overvåket WSD, burde legge følgende forutsetninger til grunn: At metoden skal kunne ekstrahere et finitt, betydningstagget korpus, og fortrinnsvis skal metoden prinsipielt være egnet for å generere slike finitte treningskorpora for hele åpen ordklasse-vokabularet i et språk. I tillegg kommer selvsagt at det betydningstaggede korpuset skal være mest mulig korrekt betydningstagget, slik også et manuelt tagget korpus vil være. Selv om vi må ta et lite forbehold siden denne oppgavens resultater rapporterer fra bare ett testlemma, synes den presenterte metoden for automatisk betydningstagging (ABT) lovende ut fra de overnevnte forutsetninger.

Ettersom metoden baserer betydningstaggingen på oversettelseskorrespondanser sortert ved speilmetoden, viser resultatene at ABT-metoden er begrenset til å kunne tagge de forekomster av testlemmaet som har en tilstrekkelig klar oversettelseskorrespondanse (det vil si en oversettelseskorrespondanse som inngår i speilmetodens betydningspartisjoner). Treningen av WSD-klassifikatorene i (kap.4) indikerer imidlertid at selv om ABT-materialet bare utgjør ca. halvparten av ENPCs forekomster av testlemmaet, later ikke selve forskjellen i treningsmengde til å påvirke klassifikatorens læringsevne i særlig grad.

Derimot antyder resultatene fra alle tre klassifikatorer at ujevn fordeling mellom mengden av treningsinstanser per betydningskategori kan utgjøre et problem med tanke på å anvende materialet som treningskorpus for en WSD-klassifikator. En ujevn fordeling mellom betydninger i et gitt korpus er i og for seg en faktor som vanskelig kan kontrolleres på forhånd når et parallellkorpus settes sammen. Likeledes kan heller ikke ABT-metoden kontrollere dette, ettersom ABT-metoden er begrenset til de instanser som foreligger i et parallellkorpus (uten å kunne supplere ytterligere treningsinstanser fra et hvilket som helst annet korpus, slik det er mulig å gjøre manuelt).

ABT-metodens mulighet for å ekstrahere treningskorpora for et størst mulig vokabular i et språk avhenger av tilgang på et ordlenket parallellkorpus. Siden ordparallellstillingen av ENPC per i dag ikke er på plass, må for det første oversettelseskorrespondanser som input til speilmetoden ekserperes manuelt. Dette innebærer i sin tur at ABT-metoden er begrenset til det vokabularet som speilmetoden manuelt har fått informasjon om. Når ordlenkingen av korpuset er fullført, kan speilmetoden derimot anvendes direkte på ENPC. Dersom det ordlenkede materialet er tilfredsstillende, betyr dette at ABT-metoden prinsipielt ville kunne anvendes på alle ord i vokabularet som speilmetoden genererer betydningsskiller for. Ordparallellstilling av korpuset vil således kunne øke ABT-metodens generaliseringspotensial. Vi vil da også ha mulighet for å kunne undersøke i hvilken grad den svært skjeve fordelingen mellom eksempelinstanser for hver betydning var særlig uheldig for denne oppgavens testlemma, eller om dette utgjør en generell svakhet ved et treningskorpus som er avhengig av de instanser som suppleres i et parallellkorpus.

Når det gjelder ABT-metodens evne til å produsere korrekte betydningstagger, indikerer de presenterte resultatene at selve ABT-metoden synes svært lovende. Det ble observert at problemene først og fremst er forbundet med de praktiske, implementeringsmessige omgivelsene. Det har blitt vist at det største problemet for ABT-metodens presisjon var mangelen på et ordlenket materiale, et problem som forsterkes i kombinasjon med svak preprosessering av ENPC. Derfor ville en tilgang på ordlenket materiale også kunne forventes å forbedre ABT-metodens presisjon. Resultatene indikerer at metoden har en presisjon på 83,3 % hvis vi også inkluderer problemer som en følge av implementeringens praktiske begrensninger. Om vi kun tar hensyn til metodisk relaterte feil, økes derimot presisjonen i ABT-materialet til 96 %. Treningen av en WSD-klassifikator i

(kap.4) synes å gjenspeile dette: De trente klassifikatorene antyder en markant forskjell i klassifikatorens prestasjon med og uten praktisk relaterte feilkilder. En klassifikator trent på et automatisk materiale med kun metodisk relaterte feil synes å være nesten på høyde med en klassifikator trent på et manuelt tagget korpus, på tross av at det manuelt taggede materialet omfattet alle nesten dobbelt så mange instanser av testlemmaet som enn ABT-materialet.

Speilmetoden later til å være en lovende ressurs for ekstrahering av et finitt, betydningstagget korpus. For det første genererer den selve betydningsskillene som en automatisk betydningstagger kan basere seg på, og for det andre utgjør speilmetodens betydningsmessige sortering av oversettelseskorrespondanser selve kunnskapskilden for ABT-metodens uovervåkede betydningstagging av instanser. Grunnet fattige oversettelsesdata i parallellkorpuset observerte vi at speilmetoden tvinges til å generere flere betydningspartisjoner enn hva som kanskje er ønskelig. Ut fra resultatene ved trening av en klassifikator på basis av disse betydningsskillene, har det imidlertid blitt argumentert for at klassifikatoren faktisk ikke først og fremst synes å ha problemer med speilmetodens betydningsskiller. Snarere synes alle tre klassifikatorenes hovedproblem å være relatert til den totale mengden treningsmateriale som ENPC supplerte for testlemmaet, og kanskje primært til den svært skjeve fordelingen mellom betydningene i korpuset.

Siden størrelsen på korpuset er et problem, ville en interessant videreutvikling av metoden kunne være å la det ekstraherte ABT-materialet utgjøre såkalte frø for en "bootstrapping"-algoritme, i tråd med Yarowskys forsøk som ble omtalt i (1.4.3). Yarowsky (1995) presenterte en "bootstrapping"-metode for WSD hvor de såkalte frøene ble generert automatisk. Basert på en klassifikator trent på disse frøene, ekstraheres iterativt et stadig større treningsmateriale. Siden denne oppgavens resultater for en klassifikator trent på ABT-materialet synes å ha et relativt høyt nivå, burde ABT-materialet kunne være anvendbare som "frø". Et interessant poeng ved Yarowskys forsøk er at implementeringen forsøker å ta høyde for hvorvidt klassifikatoren er "sikker" på at en potensiell ny treningsinstans blir korrekt inkludert i treningsmaterialet. Implementeringen muliggjør som sådan en form for "selvkorrigering": Hvis det stadig økende treningsmaterialet gir grunnlag for å "se" at en instans som ble lagt til på et tidligere stadium er galt klassifisert, kan systemet gå tilbake og korrigere feilen. Siden ABT-materialet vi har sett i denne oppgaven har svært få instanser av "matrett"-betydningen av *rett*N, ville en slik korreksjonsmulighet kunne være av stor betydning.

Gitt et ordlenket materiale som åpner for at speilmetoden kan generere mer omfattende semantisk informasjon, kunne det også være mulig "utvide" treningskorpuset med utgangspunkt i de semantiske trekkene som speilmetoden genererer for hver ordbetydning. Dette kunne gjøres ved å registrere kontekstordenes semantiske trekk i trekkvektorene framfor selve ordene. I (2.3.3) ble det for eksempel illustrert at *rett*N i den største "matrett"-betydningen (representert ved DISH som betydningstag) er assosiert med semantiske trekk som [mat1|supper2], [middag1|food5] og [måltid1|dish3] ifølge speilmetoden. Tanken er at kontekstordene til en gitt instans av *rett*N kan vise seg å dele semantiske trekk med øvrige ord i parallellkorpus, slik at disse kanskje kan grupperes sammen. Som eksempel kan vi tenke oss at et ord *a* (la oss si *mat*) i en kontekst til *rett*N deler semantiske trekk med et ord *b* (for eksempel *middag*) som finnes i korpus, men som ikke finnes i noen kontekster for *rett*N. Hvis ord *b* deler semantiske trekk med ord *a*, og *a* er assosiert med en viss betydning av *rett*N, er det ikke utenkelig at det skyldes tilfeldigheter i et sparsommelig korpus at ikke også *b* forekom i *rett*Ns kontekst. Det vil derfor være interessant å forsøke å gruppere kontekstord sammen med deres semantisk nærbeslektede ord ifra hele korpus. Hvis klassifikatoren kan bruke treningskorpus til å ekstrahere et sett av semantiske trekk (framfor et spesifikt, endelig sett av ord), vil det bety at klassifikatoren har generalisert ut over de ord som faktisk er belagt

i konteksten. Å teste ut dette forutsetter imidlertid tilgang på et ordlenket korpus, slik at speilmetoden kan generere semantiske trekk for kontekstordene til det flertydige ordet.

6. Konklusjon

Denne hovedoppgaven var rettet mot problemet med manglende tilgang på betydningstaggede treningskorpora for å kunne teste ut korpusbaserte, overvåkede metoder for WSD i større skala. Oppgaven har foreslått en metode som anvender oversettelseskorrespondanser i et parallellkorpus, sortert betydningsmessig ved speilmetoden, som bakenforliggende ressurs for å ekstrahere et finitt betydningstagget korpus. Denne metoden er prinsipielt ment for å ekstrahere finitte treningskorpora som *bakenforliggende* ressurs for trening av en klassifikator etter prinsippet for overvåket WSD. I neste steg ble det så trent en WSD-klassifikator basert på det betydningstaggede materialet.

Vi har sett at speilmetoden synes å være en velegnet ressurs for å supplere både betydningskategorier og å utgjøre selve kunnskapskilden for å utføre uovervåket betydningstagging: Resultatene fra implementeringen av ABT-metoden og treningen av en klassifikator antyder at tilnærmingen metodisk sett synes lovende. Imidlertid har vi også observert at de presenterte resultatene bærer preg av parallellkorpusets relativt sparsommelige data, både med hensyn til speilmetodens utledede betydningsskiller, størrelsen på det ekstraherte treningskorpuset og en klassifikators læringsevne. Dette er problemer vi kan karakterisere som å være av mer praktisk art, og det har blitt skissert noen mulige videreføringer av den presenterte metoden som retter seg mot å øke størrelsen på treningskorpuset.

De få svakhetene som er observert av metodisk art antyder at den presenterte metoden for automatisk betydningstagging kan ha et reelt potensial som alternativ til å manuelt ekstrahere betydningstaggede korpora. Med forbedrede korpusressurser (først og fremst at ordlenkingen av parallellkorpuset ENPC blir ferdigstilt) gir den presenterte ABT-metodens resultater dermed godt håp for videreutvikling for å anvende metoden i større skala.

Referanser

Agirre, E. & Martinez, D. (2001): "Knowledge Sources for Word Sense Disambiguation". I: Matousek, V., Mautner, P., Moucek, R., Tauser, K. (eds.) (2001) *Proceedings of the Fourth International Conference TSD*. Plzen (Pilsen), Czech Republic, September 2001. Springer Verlag Lecture Notes in Computer Science series.

Arnold, D., Lorna Balkan, Siety Meijer, R. Lee Humphreys & L. Sadler (2000/1994): "A bit of history". Kap 1.4 I: *Machine Translation. An Introductory Guide.* NCC Blackwell, London.

Bondi Johannesen, J. (1998): *En grammatisk tagger for norsk (bokmål)*. (Tekstlaboratoriet 1998). Tilgjengelig på: http://www.hf.uio.no/tekstlab/tagger2.html

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. & Mercer, R. L. (1991): "Word Sense Disambiguation using Statistical Methods". I: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, New Mexico.
Tilgjengelig på: http://acl.ldc.upenn.edu/P/P91/P91-1034.pdf

Chodorow, M., Leacock, C. & Miller, G.A. (2000): "*A Topical/Local Classifier for Word Sense Disambiguation"*.. s.115-120 i: *Computers and the Humanities*. *Special Issue on SENSEVAL* 34: 115-120. Kluwer Academic Publishers.

Chodorow, M., Leacock, C. & Miller, G.A. (1998): "Using corpus statistics and WordNet relations for sense identification". s.147-165 i: *Computational Linguistics*.

Daelemans, W., van der Sloot, K. & Zavrel, J. (2001): "TiMBL: Tilburg Memory-Based Learner. Version 4.1. Reference Guide". *ILK Technical Report – ILK 01-04*.
Tilgjengelig på: http://ilk.kub.nl/downloads/pub/papers/ilk0104.ps.gz.

Daelemans, W., van den Bosch, A. & Zavrel, J. (1999): "Forgetting Exceptions is Harmful in Language Learning". s 11-43 i: *Machine Learning, special issue in natural language learning.*

Diab, M. & Resnik, P. (2002): "An Unsupervised Method for Word Sense Tagging using Parallel Corpora". I: *40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, July.

Dini, L., Di Tomaso, V. & Segond, F. (2000): "GINGERII: An Example-Driven Word Sense Disambiguator". s. 121-126 i: *Computers and the Humanities*. *Special Issue on SENSEVAL* 34: 121-126. Kluwer Academic Publishers.

Dyvik, H. (2002): "Translations as Semantic Mirrors: From Parallel Corpus to Wordnet". *ICAME (International Computer Archive of Modern and Medieval English) Conference 2002*. Göteborg, Sverige, mai 2002.
Tilgjengelig på: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorpapers.html

Dyvik, H. (1998a): "A Translational Basis for Semantics". s. 51-86 i: Stig Johansson and Signe Oksefjell (eds.)(1998): *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*. Rodopi.
Tilgjengelig på: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorpapers.html

Dyvik, H. (1998b): "Translations as Semantic Mirrors". s. 24-44 i: *Proceedings of Workshop W13: Multilinguality in the LLExicon II, The 13th Biennial European Conference on Artificial Intelligence (ECAI 98)*. Brighton, UK. 24-44.

Escudero,G., Màrquez, L. & Rigau, G. (2000): "A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation". s 31-36 i: *Proceedings of CoNLL-2000 and LLL-2000*. Lisbon, Portugal.

Gale, W., Church, K. & Yarowsky, D. (1992): "A method for disambiguating word senses in a large corpus". Kap.26 i: *Computers and the Humanities*. (1992).

Hoste, V., Hendrix, I., Daelemans, W. & van der Bosch, A. (2001): "Parameter Optimization for Machine-Learning of Word Sense Disambiguation". I: *Natural Language Engineering*. UK.

Ide, N., Erjavec, T. & Tufis, D. (2002): "Sense Discrimination with Parallel Corpora". s. 54-60 i: *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*: *Recent Successes and Future Directions*. Philadelphia, July 2002. Association for Computational Linguistics.
Tilgjengelig på: http://acl.ldc.upenn.edu/acl2002/WSD/pdfs/WSD008.pdf

Ide, N. (1999): "Word Sense Disambiguation Using Cross-Lingual Information". I: *Proceedings of ACH-ALLC '99 International Humanities Computing Conference*. Charlottesville, Virginia.
Tilgjengelig på: http://jefferson.village.virginia.edu/ach-allc.99/proceedings.

Ide, N. & Veronis, J. (1998): "Word Sense Disambiguation: The State of the Art". s. 1-40 i: *Computational Linguistics.* Tilgjengelig på: http://www.up.univ.mrs.fr/~veronis/pdf/1998wsd.pdf

Johansson, S., Ebeling, J. & Oksefjell, S. (1999/2002): *English-Norwegian Parallel Corpus: Manual.* Department of British and American Studies, University of Oslo.
Tilgjengelig på: http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html.

Jurafsky, D. & Martin, J.H. (2000): "Word Sense Disambiguation and Information Retrieval". s. 631-647 i: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Krovetz, R. (1998): "More than one sense per discourse". I: NEC Princeton NJ Labs., Research Memorandum.

Leacock, C. & Chodorow, M. (1998): "Combining Local Context and WordNet Similarity for Word Sense Identification". Kap.11 i: Fellbaum, C. (red.) (1998): *WordNet. An Electronic Lexical Databse*. MIT Press.

Miller, G. A. (1998): "Nouns in WordNet". Kap.1 i: Fellbaum, C. (red.) (1998): *WordNet. An Electronic Lexical Databse*. MIT Press.

Moon, R. (2000): "Lexicography and Disambiguation: The size of the Problem". s 99-102 i: *Computers and the Humanities. Special Issue on SENSEVAL.* Kluwer Academic Publishers.

Och, F. J. & Ney, H. (2000): "Improved statistical alignment models". I: *Proceedings of 37th Annual Meeting of the Association for Computational Lingustics (ACL'00)*. Hong Kong, October.

Resnik, P. (1995): "Disambiguating Noun Groupings with respect to WordNet Senses". I: *Proceedings of the 3dr Workshop on Very Large Corpora.* MIT.

Resnik, P. & Yarowsky, D. (2000): "Distiguishing Systems and Distiguishing Senses: New Evaluation Methods for WSD". s. 113-133 i: *Natural Language Engineering*. Cambridge University Press.
Tilgjengelig på: http://www.cs.jhu.edu/~Yarowsky/pubs.htm.

Resnik, P. & Yarowsky, D. (1997): "A Perspective on WSD Methods and Their Evaluation". s. 70-86 i: *ACL SIGLEX Workshop om Tagging Text with Lexical Semantics: Why, What and How?* April, Washington, D.C.

Saeed, J. I. (1997): *Semantics*. Black Publishers Ltd.

Santorini, B. (1991): *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*.
Tilgjengelig på: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-tagset.ps

Stevenson, M. (2003): *Word Sense Disambiguation - The Case for Combinations of Knowledge Sources*. CSLI Publications.

Stevenson, M. & Wilks, Y. (2001): "The interaction of knowledge sources in Word Sense Disambiguation". s. 321-349 i: *Computational Linguistics*.

Stevenson, M., Cunningham, H., Wilks, Y. (1998): "Sense Tagging and Language Engineering". s. 185-189 i: *Proceedings of the 8th European Conference on Artificial Intellingence.(ECAI-98).* Brighton, UK.

Thunes, M. (2003): "Ekserpering av oversettelseskorrespondanser fra parallelltekst".
Tilgjengelig fra: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorpapers.html

Veenstra, J, van den Bosch, A., Buchholz, S., Daelemans, W. & Zavrel, J. (2000): "Memory-Based Word Sense Disambiguation". s. 171-177 i: *Computers and the Humanities*. *Special Issue on SENSEVAL*. Kluwer Academic Publishers.

Weaver, W. (1949): "Translation. Mimeographed, 12pp., July 15, 1949" Reprinted in Locke, W.N. & Botth, A.D.(1955) (eds), *Machine translation of Languages.* John Wiley & Sons, New York.

Wilks, Y. (2000): "Is Word Sense Disambiguation Just One More NLP Task?". s. 235-243 i: *Computers and the Humanities*. *Special Issue on SENSEVAL.* Kluwer Academic Publishers.

Wilks, Y. & Stevenson, M. (1998): "Word Sense Disambiguation Using Optimised Combinations of Knowledge Sources". s. 1398-1402 i: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics.(COLING-ACL '98)*. Montreal, Canada.

Wilks, Y. & Stevenson, M. (1997): " Combining Independent Knowledge Sources for Word Sense Disambiguation". s. 1-7 i: *Proceedings of the Conference Recent Advances in Natural Language Processing*. Tzigov Chark, Bulgaria.

Wilks, Y. & Stevenson, M. (1996): "The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?". I: *Sheffield Department of Computer Science, Research Memoranda*, CS-96-05. Tilgjengelig på: http://www.dcs.shef.ac.uk/~yorick/papers.html

Yarowsky, D. (1995): "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods". s. 189-196 i: *Proceedings of the 33rd Annual Meeting of the Association for Computatioanal Linguistics*. Cambridge, MA. http://www.cs.jhu.edu/~Yarowsky/pubs.htm.

Yarowsky, D. (1993): "One sense per Collocation". s. 266-271 i: *Proceedings ARPA Human Language Technology Workshop*. Princeton, N.J.

Yarowsky, D. (1992): "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora". I: *Proceedings of COLING92*.

Appendiks 1: LISP-implementering av automatisk betydningstagger

```
;;;   -*- Mode: LISP; Package: CGP; BASE: 10; Syntax: ANSI-Common-Lisp; -*-
;;
;; Version 0.1
;;
;; IMPLEMENTATION OF AN AUTOMATIC SENSE-TAGGER,
;; USING SENSE PARTITIONS FROM THE METHOD "SEMANTIC MIRROS" AS KNOWLEDGE SOURCE

(in-package "CGP")
(cl-user::mk :semantic-mirrors-gunn)
(cl-user::mk :xml)

;;-----------------------------------------------------------------------------------------------------
;; Basic helping methods to retrieve the one or more pos-tags/lemmas of a token:
;;-----------------------------------------------------------------------------------------------------

;; retrieves all the pos-tags and lemmas of a token
(defmethod lemmas-and-pos-tags (token)
  (rest token))

;; makes a list of only the pos-tags (excluding the lemma) of a token
(defmethod all-pos-tags (token)
  (if token
      (if (listp (first token))
              (cons (rest (first token))
                          (all-pos-tags (rest token)))
            (all-pos-tags (rest token)))))
#+test
(print (all-pos-tags '("retten" ("rett" . SUBST)("rett" . ADJ))))

;; makes a list of only the lemmas (excluding pos-tags) of a token
(defmethod all-token-lemmas (token)
  (if token
      (if (listp (first token))
              (cons (first (first token))
                          (all-token-lemmas (rest token)))
            (all-token-lemmas (rest token)))))
#+test
(print (all-token-lemmas  '("retten" ("rette" . SUBST) ("rett" . SUBST))))

;;-----------------------------------------------------------------------------------------------------
;; helping functions/methods for the sense-tagger:
;;-----------------------------------------------------------------------------------------------------

;; compares a given L2lemma against all sense-partitions. Only returns a match if the L2lemma has the desired pos-tag
(defun sense-tag (L2lemma L2pos L2desired-pos sense-partitions)
  (if (null sense-partitions) nil
    (if (and (intersection L2lemma
                                (first sense-partitions) :test #'string-equal)
              (equal (union L2pos L2desired-pos) L2desired-pos))
        (first (first sense-partitions))
      (sense-tag L2lemma L2pos L2desired-pos (rest sense-partitions)))))

;; calculates a word/lemma's sentence position
(defmethod sentence-position (( sentence compressed-prealigned-sentence )
                                        ( token-num integer))
  (/ token-num (length (sentence-array sentence))))
```

```
;; traverses the corresponding sentence and returns all potential sense-tags along with the sentence position of each
(defmethod potential-sense-tags ((corresponding-sentences list)
                                 (sense-partitions list)
                                         (L2desired-pos list))
  (let ((found-tags ""))
    (map nil (lambda (sentence)
                 (loop
            for token across (sentence-array sentence)
            for L2token-num from 0
            do
            (let ((L2lemma (remove-duplicates (all-token-lemmas token)))
                  (L2pos (remove-duplicates (all-pos-tags token))))
            (when
                 (stringp (first L2lemma))
              (if (sense-tag L2lemma L2pos L2desired-pos sense-partitions)
                  (setf found-tags
                        (cons (cons (sense-tag L2lemma
                              L2pos
                              L2desired-pos
                              sense-partitions)
                                              (sentence-position sentence L2token-num))
                         found-tags))
                nil)))))
             corresponding-sentences)
    (reverse found-tags)))

;; Returns the sense-tag whose L2-sentence-position is the closest, relatively to the position of the search-lemma in L1
(defun closest-sem-tag (L1lemma-position potential-sensetag-positions)
  (let ((shortest-distance nil)
        (closest-sem-tag nil))
    (mapcar (lambda (target-number)
                (let ((distance (abs (- L1lemma-position (rest target-number)))))
                      (when (or (not shortest-distance)
                                (< distance shortest-distance))
                        (setf shortest-distance distance
                              closest-sem-tag (first target-number)))))
              potential-sensetag-positions)
    closest-sem-tag))

#+test
(print (closest-sem-tag '7/55
                  '(("claim" . 3/22) ("court" . 7/33))))

;; Checks if there were more than one potential sense-tags. If so, returns the closest semantic tag
;; (using the function closest-sem-tag above)
(defun choose-sense-tag (sense-tags sentence-pos)
  (if (> (length sense-tags) 1)
      (setf sense-tags (closest-sem-tag sentence-pos sense-tags))
    (setf sense-tags (first (first sense-tags)))))
```

```
;;-------------------------------------------------------------------------------------
;; main function which sense-tags all the possible instances in corpus:
;; (the output formats (print-local-context etc.) are documented in appendix 2)
;;-------------------------------------------------------------------------------------

(defun sense-tag-instances ( search-lemma-list sense-partitions pos L2desired-pos)
  (let ((number-of-sentences-total 0)(tagged 0)(untagged 0))
    (corpus::with-documents (document name *enpc*)
      (map nil (lambda (sentence)
            (setf *sentence* sentence)
            (loop
               for token across (sentence-array sentence)
               for L1token-num from 0
               do
                 (when (and (intersection search-lemma-list (all-token-lemmas token) :test #'string-equal)
                                    (member pos (all-pos-tags token)))
                        (let ((sentence-pos (sentence-position sentence L1token-num))
                          (potential-sense-tags (potential-sense-tags (corresponding-sentences sentence)
                                                sense-partitions
                                                L2desired-pos)))
                          (if (listp potential-sense-tags)
                              (and (incf tagged)
                           (format t "~a~a~%"
                                                ;;local context:
                              (print-local-context sentence search-lemma-list L1token-num context-window)
                              ;;the n nearest keywords:
                              ;;(csv-formatting
                                                ;;(collect-keywords sentence L1token-num keyword-window))
                              ;;The n nearest characteristic keywords (based on relative frequency):
                                        ;;(csv-formatting
                                                ;;(characteristic-keyword-context sentence L1token-num keyword-window docu))
                              (choose-sense-tag potential-sense-tags sentence-pos)))
                                   (incf untagged))
                          (incf number-of-sentences-total)))))
           (document-sentences document)))
    (format t "~% Number of sentences total: ~a ~% Number of sentences tagged: ~a ~% Number of sentences untagged: ~a ~%"
               number-of-sentences-total tagged untagged)))

#+test
(sense-tag-instances '("rett")
                  '(("course")
                    ("court" "justification")
                    ("claim" "entitlement" "law" "option" "order" "right")
                    ("dish" "food" "special" "supper"))
                   'SUBST
                  '(NN NNS))
```

```
;;-------------------------------------------------------------------------------------
;; Optimization attempt:
;; Attempt to only produce a sense-tag if the position of the sense-tag is closer
;; than e.g. 1/3, relatively to the position of the lemma to be tagged.
;;-------------------------------------------------------------------------------------

;;---------------------------------------
;; helping functions/methods:
;;---------------------------------------

#+less-good-threshold
(defun close-enough? (source-number target-number)
  (< (abs (- source-number target-number)) 1/3))

(defun close-enough? (source-number target-number)
  (< (abs (- source-number target-number)) 1/4))

;;redefined from original function above
(defun closest-sem-tag2 (L1lemma-position potential-sensetag-positions)
  (let ((shortest-distance nil)
        (closest-sem-tag nil))
    (mapcar (lambda (target-number)
                (let ((distance (abs (- L1lemma-position (rest target-number)))))
                        (when (or (not shortest-distance)
                                        (< distance shortest-distance))
                          (setf shortest-distance distance
                                        ;;closest-sem-tag (first target-number)))))
                                        ;;now replaced by:
                                        closest-sem-tag target-number))))
              potential-sensetag-positions)
    closest-sem-tag))

;; redefined from above
(defun choose-sense-tag2 (sense-tags sentence-pos)
  (if (= (length sense-tags) 1)
      (when (close-enough? sentence-pos
                        (rest (first sense-tags)))
          (first (first sense-tags)))
    (when (close-enough? sentence-pos
                      (rest (closest-sem-tag2 sentence-pos sense-tags)))
        (first (closest-sem-tag2 sentence-pos sense-tags)))))
#+test
(print (choose-sense-tag2 '(("claim" . 3/22) ("dish" . 13/22)) '5/23))
```

```
;;-------------------------------------------------------------------------------------
;; main function for optimization attempt (redefined from sense-tag-instances above)
;;-------------------------------------------------------------------------------------

 (defun sensetag-close-enough ( search-lemma-list sense-partitions pos L2desired-pos)
  (let ((number-of-sentences-total 0)(tagged 0)(untagged 0))
    (corpus::with-documents (document name *enpc*)
      (map nil (lambda (sentence)
                       (setf *sentence* sentence)
              (loop
                 for token across (sentence-array sentence)
                 for L1token-num from 0
                 do
                   (when (and (intersection search-lemma-list (all-token-lemmas token) :test #'string-equal)
                                        (member pos (all-pos-tags token)))
                                 (let ((sentence-pos (sentence-position sentence L1token-num))
                           (potential-sense-tags   (potential-sense-tags (corresponding-sentences sentence)
                                                    sense-partitions
                                                    L2desired-pos)))
                           (when (listp potential-sense-tags)
                              (if (stringp (choose-sense-tag2 potential-sense-tags sentence-pos))
                                  (and (incf tagged)
                                   (format t "~a~a  ~%"
                                   (print-local-context sentence search-lemma-list L1token-num context-window)
                                   (choose-sense-tag2 potential-sense-tags sentence-pos)))
                                    (incf untagged)))
                                 (incf number-of-sentences-total)))))
              (document-sentences document)))
   (format t "~% Number of sentences total: ~a ~% Number of sentences tagged: ~a ~% Number of sentences untagged: ~a ~%"
                 number-of-sentences-total tagged untagged)))

#+test
(sensetag-close-enough '("rett")
                       '(("course")
                        ("court" "justification")
                                 ("claim" "entitlement" "law" "option" "order" "right")
                        ("dish" "food" "special" "supper"))
                                'SUBST
                       '(NN NNS))
```

Appendiks 2: LISP-implementering av å ekstrahere kontekst for målordet

```
;;;   -*- Mode: LISP; Package: CGP; BASE: 10; Syntax: ANSI-Common-Lisp; -*-
;;
;; Version 0.1
;;
;; IMPLEMENTATION OF THE EXTRACTION OF CONTEXT AROUND THE SENSE-TAGGED WORD
;; WHICH IS APPLIED IN THE MAIN FUNCTION SENSE-TAG-INSTANCES

(in-package "CGP")

;;----------------
;; local context
;;----------------

;;sets a default value for the position-specific, local context on each side of the tagged word
(defparameter context-window 2)

;; collects the lemma of a wanted sentence position n
(defmethod context-token ((sentence compressed-prealigned-sentence)
                          (search-lemma list)
                          (token-num integer) ;; which sentential position the ambiguous word has
                                  (context-offset integer)) ;; context-word number n to fill the context-window
   (let ((sentence-num (+ token-num context-offset)))
      (cond ((null sentence) nil)
            ((= sentence-num token-num) (first search-lemma))
            ((< sentence-num 1) nil) ;; i.e. if not enough left-side words
            ((> sentence-num (1- (length (sentence-array sentence)))) nil) ;; i.e. not enough right-side words
            ;; Builds LEMMA-context; note the risk that some occurrences may have been falsely lemmatized:
            (t (token-lemma (aref (sentence-array sentence) sentence-num))))))
#+test
(print (context-token *sentence* '("rett") 4 -2))

;; builds the local context in csv-format (comma-separated values
(defmethod print-local-context ((sentence compressed-prealigned-sentence)
                                (search-lemma list)
                                (token-num integer) ;; the sentential position of the ambiguous word
                                (context-window integer))
   (let ((context ""))
      (loop
       for context-offset from (- context-window) to context-window
       do (setf context (concatenate 'string context
                          (if (wordp (context-token sentence search-lemma token-num context-offset))
                          (context-token sentence search-lemma token-num context-offset) "-" )
                        ",")))
   context))

#+test
(princ (print-local-context *sentence* '("rett") 4 6))
```

```
;;-------------------------------------------------------
;; n nearest keywords in a larger context
;; -------------------------------------------------------

;; basic helping functions/methods to extract keywords

;; sets a default value for the number of nearest keywords to be collected:
(defparameter keyword-window 12)

;;Based on ENPC's pos-tags, returns T if a token is NOT a grammatical word
(defun open-classp (token)
  (let ((pos (all-pos-tags token)))
    (and (wordp token)
         (not (or  (member 'DET pos)
                   (member 'PREP pos)
             (member 'PRON pos)
             (member 'SBU pos)
             (member 'INTERJ pos)
             (member 'KONJ pos)
             (member 'INF-MERKE pos)
             (equal (token-lemma token) "NIL")
             (equal (first token) "NIL"))))))

#+test
(print (open-classp '("på" ("på" . PREP)))) ;; returns NIL
(print (open-classp '("glade" ("glad" . ADJ)))) ;; returns T

;; Returns the token in the wanted sentential position, may traverse sentence limits to find the needed token.
;; Returns nil when there is no next/previous sentence to fetch a token from.
(defparameter kw '())
(defmethod context-keyword ((sentence compressed-prealigned-sentence)
                            (token-num integer) ;; sentential position of the sense-tagged word
                            (context-offset integer)) ;; context-word number n to fill the context-window
  (let ((sentence-num (+ token-num context-offset)))
    (cond ((null sentence) nil)
          ((< sentence-num 1) ;; collecting from the left-side of the sense-tagged word
           (if (null (prev-sentence sentence)) nil
             (context-keyword (prev-sentence sentence)
                              ;; -1 because there is an initial zero element in each sentence array:
                              (- (length (sentence-array (prev-sentence sentence))) 1)
                              (+ token-num context-offset))))
           ;; -1 because of the initial zero element of a sentence array:
          ((> sentence-num (- (length (sentence-array sentence)) 1)) ;;collecting from the right-side of the sense-tagged word
            (if (null (next-sentence sentence)) nil
              (context-keyword (next-sentence sentence)
                               0
                               ;; -1 because of the initial zero element of a sentence array:
                               (- (+ token-num context-offset)(- (length (sentence-array sentence)) 1)))))
           (t (aref (sentence-array sentence) sentence-num)))))

#+test
(print (context-keyword *sentence*  4 -5))

;; Prints the output format of lemma/word+comma (csv-format; comma-separated values):
(defmethod csv-formatting ((tokens list))
  (let ((string ""))
    (mapcar (lambda (token)
              (setf string (concatenate 'string string (if token (caadr token) "-") ",")))
            tokens)
    string))
```

```
;; ---------------------------------------------------------------------
;; Keyword definition 1
;;     weeds out grammatical words (based on ENPCs pos-tags) and
;;     the 10,000 most frequent words in  Norwegian (based on Tekstlaboratoriets frekvensordliste)
;; ---------------------------------------------------------------------

;; helping methods/functions for keyword definition 1

;; Returns T if the word is (assumed to be) a valid contextual keyword:
;; NOT a grammtical word and NOT a member of Tekstlaboratoriets Frekvensordliste
(defun token-keywordp (token)
   (and (open-classp token)
            (not (gethash (first token) *frekvensordliste*))))

;; collects keywords from the left-side of the sense-tagged word
(defun n-previous-keywords (keyword-window sentence token-num context-offset)
  (when (> keyword-window 0)
    (decf context-offset) ;;keeps track of the counting backwards through sentences, token by token
    (let ((token (context-keyword sentence token-num context-offset)))
      (cond ((null token) kw) ;; then there are not enough prev-sentences in the document
            ((and (token-keywordp token)
                  (not (string-member (token-lemma token) (mapcar 'token-lemma kw))));; avoiding duplicates
             (and (setf kw (append kw (list token)))
                  (n-previous-keywords (decf keyword-window) sentence token-num context-offset)))
            (t (n-previous-keywords keyword-window sentence token-num context-offset))))) kw)

;; collects keywords from the right-side of the sense-tagged word
(defun n-next-keywords (keyword-window sentence token-num context-offset)
  (when (> keyword-window 0)
    (incf context-offset) ;; counts forwards through the sentence
    (let ((token (context-keyword sentence token-num context-offset)))
      (cond ((null token) kw) ;; then there are not enough next-sentences in the document
            ((and (token-keywordp token)
                  (not (string-member (token-lemma token) (mapcar 'token-lemma kw))))
             (and (setf kw (append kw (list token)))
                  (n-next-keywords (decf keyword-window) sentence token-num context-offset)))
            (t (n-next-keywords keyword-window sentence token-num context-offset))))) kw)

#+test
(print (n-previous-keywords 6 *sentence* 4 0))
#+test
(print (n-next-keywords 6 *sentence* 4 0))


;; main method

;; checks if there were not enough prev- or next-sentences to fill an n/2-window. If there were not sufficiently many
;; words on one of the sides, the rest is extracted from the other side of the context window.
(defmethod collect-keywords ((sentence compressed-prealigned-sentence) (token-num integer) (keyword-window integer))
  (setf kw ())
  (let ((keyword-win (/ keyword-window 2))
        (prev-keyw (n-previous-keywords (/ keyword-window 2) sentence token-num 0))
        (next-keyw (n-next-keywords (/ keyword-window 2) sentence token-num 0)))
    (cond ((< (length prev-keyw) keyword-win) ;; then there were not enough words from left-side context
           (and (setf (kw ())
                (n-previous-keywords keyword-win sentence token-num 0)
                (n-next-keywords (+ keyword-win (- keyword-win (length prev-keyw)))
                      sentence
                      token-num
                      0)))
          ((< (length next-keyw) keyword-window) ;; then there were not enough words from right-side context
           (and (setf kw ())
                (n-previous-keywords  (+ keyword-win (- keyword-window (length next-keyw)))
                                      sentence
                                      token-num
                                      0)
                (n-next-keywords keyword-win sentence token-num 0)))
           (t kw))))

#+test
(princ (csv-formatting (collect-keywords *sentence* 4 keyword-window)))
```

```
;; ------------------------------------------------------------------------
;; Keyword definition 2
;;     collects the "n nearest characteristic keywords" of the document,
;;     weeding out closed-class (grammatical) words.  A word is "characteristic"
;;     for a  document if it is more frequent within the document than in the whole ENPC
;; ------------------------------------------------------------------------

;; helping methods/functions for keyword definition 2

;; Total number of words in ENPC corpus according to web-page
;; http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html#_Toc445194135 is:
;; Original texts Norwegian: 629900 + Translated texts Norwegian:  661 500 = 1,291400
(defparameter enpc-size-norw 1291400)

;; updates statistics, method implemented by Sindre Sørensen (the method definition is not supplied here)
(document-statistics-update *enpc*)

;; method for counting the frequency of a lemma implemented by Sindre Sørensen (its definition is not supplied here)
#+test
(print (frequency "rett" docu))
#+test
(print (frequency "rett" *enpc*))

;; counts number of words in a document
(defmethod word-count ((doc corpus::document))
  (let ((count 0))
    (loop for sentence across (document-sentences doc)
          do (map nil (lambda (token)
                        (when (wordp token)
                          (incf count)))
                  (sentence-array sentence)))
    count))

;;returns T if a token is more frequent in the document than in ENPC, and is pos-tagged as an open-class token
(defmethod characteristicp ((token list) (doc corpus::document))
  (let ((doc-size (word-count doc)))
    (when (and (wordp token)
               (open-classp token)
               (not (or (equal (first token) "NIL")
                        (equal (token-lemma token) "NIL"))))
      (let ((doc-rel-frequency (/ (frequency (token-lemma token) doc) doc-size))
            (enpc-rel-frequency (/ (frequency (token-lemma token) *enpc*) enpc-size-norw)))
        (> doc-rel-frequency enpc-rel-frequency)))))

;; collects characteristic keywords from the left-side of the sense-tagged word
(defun characteristic-previous-keywords (keyword-window sentence token-num context-offset document)
  (when (> keyword-window 0)
    (decf context-offset) ;; keeps track of the counting backwards through sentences, token by token
    (let ((token (context-keyword sentence token-num context-offset)))
      (cond ((null token) kw) ;; then there are not enough prev-sentences
            ((and (characteristicp token document)
                  (not (string-member (token-lemma token) (mapcar 'token-lemma kw))));; avoiding duplicates
             (and (setf kw (append kw (list token)))
                  (characteristic-previous-keywords (decf keyword-window) sentence token-num context-offset document)))
            (t (characteristic-previous-keywords keyword-window sentence token-num context-offset document)))))
  kw)

;; collects characteristic keywords from the right-side of the sense-tagged word
(defun characteristic-next-keywords (keyword-window sentence token-num context-offset document)
  (when (> keyword-window 0)
    (incf context-offset)
    (let ((token (context-keyword sentence token-num context-offset)))
      (cond ((null token) kw) ;; then there are not enough next-sentences
            ((and (characteristicp token document)
                  (not (string-member (token-lemma token) (mapcar 'token-lemma kw))))
             (and (setf kw (append kw (list token)))
                  (characteristic-next-keywords (decf keyword-window) sentence token-num context-offset document)))
            (t (characteristic-next-keywords keyword-window sentence token-num context-offset document)))))
  kw)
#+test
(print (characteristic-previous-keywords 6 *sentence* 4 0 docu))
#+test
(print (characteristic-next-keywords 6 *sentence* 4 0 docu))
```

```
;; main method
;; checks if there were not enough prev- or next-sentences to fill an n/2-window. If there were not sufficiently many
;; words on one of the sides, the rest is extracted from the other side of the context window. (Needs one more argument
than ;; the corresponding method for keyword-definition 1)
 (defmethod characteristic-keyword-context ((sentence compressed-prealigned-sentence)
                                            (token-num integer)
                                            (keyword-window integer)
                                            (document corpus::document))
  (setf kw ())
  (let ((keyword-win (/ keyword-window 2))
        (prev-keyw (characteristic-previous-keywords (/ keyword-window 2) sentence token-num 0 document))
        (next-keyw (characteristic-next-keywords (/ keyword-window 2) sentence token-num 0 document)))
    (cond ((< (length prev-keyw) keyword-win)
           (and (setf kw ())
                (characteristic-previous-keywords keyword-win sentence token-num 0 document)
                (characteristic-next-keywords (+ keyword-win (- keyword-win (length prev-keyw)))
                                              sentence
                                              token-num
                                              0
                                              document)))
          ((< (length next-keyw) keyword-window)
           (and (setf kw ())
                (characteristic-previous-keywords  (+ keyword-win (- keyword-window (length next-keyw)))
                                                    sentence
                                                    token-num
                                                    0
                                                    document)
                (characteristic-next-keywords keyword-win sentence token-num 0 document)
           (t kw))))
#+test
(princ (csv-formatting (characteristic-keyword-context *sentence* 4 keyword-window docu)))
```

Appendiks 3: Klassifikatoren K1s delresultater. *Confusion matrix*

partisjon 1		Antall korrekte: 21/23 Prosentvis presisjon: 91,30 %		
	claim	**court**	**course**	**dish**
claim	19	1	0	0
court	0	2	0	0
course	0	0	0	0
dish	1	0	0	0

partisjon 2		Antall korrekte: 22/23 Prosentvis presisjon: 95,65%		
	claim	**court**	**course**	**dish**
claim	21	0	0	0
court	0	1	0	0
course	1	0	0	0
dish	0	0	0	0

partisjon 3		Antall korrekte: 18/23 Prosentvis presisjon: 78,26%		
	claim	**court**	**course**	**dish**
claim	15	0	1	1
court	3	2	0	0
course	0	0	0	0
dish	0	0	0	1

partisjon 4		Antall korrekte: 14/23 Prosentvis presisjon: 60,87 %		
	claim	**court**	**course**	**dish**
claim	12	2	0	0
court	3	2	0	0
course	3	0	0	0
dish	1	0	0	0

partisjon 5		Antall korrekte: 18/23 Prosentvis presisjon: 78,26%		
	claim	**court**	**course**	**dish**
claim	15	2	0	1
court	1	2	0	0
course	1	0	0	0
dish	0	0	0	1

partisjon 6		Antall korrekte: 8/23 Prosentvis presisjon: 34,78 %		
	claim	**court**	**course**	**dish**
claim	2	0	0	0
court	14	6	1	0
course	0	0	0	0
dish	0	0	0	0

partisjon 7	Antall korrekte: 18/23 Prosentvis presisjon: 78,26 %			
	claim	**court**	**course**	**dish**
claim	16	2	0	0
court	1	1	0	0
course	0	2	0	0
dish	0	0	0	1

partisjon 8	Antall korrekte: 20/23 Prosentvis presisjon: 86,96 %			
	claim	**court**	**course**	**dish**
claim	16	0	0	0
court	1	2	0	0
course	1	0	0	0
dish	1	0	0	2

partisjon 9	Antall korrekte: 21/23 Prosentvis presisjon: 91,30 %			
	claim	**court**	**course**	**dish**
claim	21	1	0	0
court	0	0	0	0
course	1	0	0	0
dish	0	0	0	0

partisjon 10	Antall korrekte: 18/21 Prosentvis presisjon: 85,71 %			
	claim	**court**	**course**	**dish**
claim	17	0	1	1
court	0	1	0	0
course	0	1	0	0
dish	0	0	0	0

Appendiks 4: Klassifikatoren K2s delresultater. *Confusion matrix*

partisjon 1		Antall korrekte: 19/20 Prosentvis presisjon: 95,00 %		
	claim	**court**	**dish**	**course**
claim	17	1	0	0
court	0	2	0	0
dish	0	0	0	0
course	0	0	0	0

partisjon 2		Antall korrekte: 20/20 Prosentvis presisjon: 100 %		
	claim	**court**	**course**	**dish**
claim	19	0	0	0
court	0	1	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 3		Antall korrekte: 17/20 Prosentvis presisjon: 85 %		
	claim	**court**	**course**	**dish**
claim	13	0	1	0
court	2	3	0	0
course	0	0	1	0
dish	0	0	0	0

partisjon 4		Antall korrekte: 15/20 Prosentvis presisjon: 75 %		
	claim	**court**	**course**	**dish**
claim	12	2	0	0
court	2	3	0	0
course	0	0	0	0
dish	1	0	0	0

partisjon 5		Antall korrekte: 14/20 Prosentvis presisjon: 70 %		
	claim	**court**	**course**	**dish**
claim	10	3	0	0
court	3	3	0	0
course	0	0	1	0
dish	0	0	0	0

partisjon 6		Antall korrekte: 11/20 Prosentvis presisjon: 55 %		
	claim	**court**	**course**	**dish**
claim	1	0	0	0
court	9	10	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 7		Antall korrekte: 17/20 Prosentvis presisjon: 85 %		
	claim	**court**	**course**	**dish**
claim	16	2	0	0
court	0	0	0	0
course	0	0	1	0
dish	0	1	0	0

partisjon 8		Antall korrekte: 17/20 Prosentvis presisjon: 85 %		
	claim	**court**	**course**	**dish**
claim	13	1	0	0
court	0	2	0	0
course	1	0	2	0
dish	1	0	0	0

partisjon 9		Antall korrekte: 20/20 Prosentvis presisjon: 100 %		
	claim	**court**	**course**	**dish**
claim	20	0	0	0
court	0	0	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 10		Antall korrekte: 18/18 Prosentvis presisjon: 100 %		
	claim	**court**	**course**	**dish**
claim	17	0	0	0
court	0	1	0	0
course	0	0	0	0
dish	0	0	0	0

Appendiks 5: Klassifikatoren K3s delresultater. *Confusion matrix*

partisjon 1		Antall korrekte: 37/40 Prosentvis presisjon: 92,5 %		
	claim	**court**	**dish**	**course**
claim	36	0	0	0
court	3	1	0	0
dish	0	0	0	0
course	0	0	0	0

partisjon 2		Antall korrekte: 37/40 Prosentvis presisjon: 92,5 %		
	claim	**court**	**course**	**dish**
claim	30	0	1	0
court	1	6	0	0
course	0	0	1	0
dish	1	0	0	0

partisjon 3		Antall korrekte: 17/40 Prosentvis presisjon: 42,5 %		
	claim	**court**	**course**	**dish**
claim	10	1	0	0
court	20	7	0	1
course	1	0	0	0
dish	0	0	0	0

partisjon 4		Antall korrekte: 33/40 Prosentvis presisjon: 82,5 %		
	claim	**court**	**course**	**dish**
claim	30	1	0	0
court	0	2	0	0
course	3	0	1	0
dish	2	1	0	0

partisjon 5		Antall korrekte: 40/40 Prosentvis presisjon: 100 %		
	claim	**court**	**course**	**dish**
claim	38	0	0	0
court	0	2	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 6		Antall korrekte: 37/40 Prosentvis presisjon: 92,5 %		
	claim	**court**	**course**	**dish**
claim	35	2	0	0
court	1	2	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 7		Antall korrekte: 38/40 Prosentvis presisjon: 95 %		
	claim	**court**	**course**	**dish**
claim	36	2	0	0
court	0	2	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 8		Antall korrekte: 36/40 Prosentvis presisjon: 90 %		
	claim	**court**	**course**	**dish**
claim	35	2	0	0
court	2	1	0	0
course	0	0	0	0
dish	0	0	0	0

partisjon 9		Antall korrekte: 38/40 Prosentvis presisjon: 95 %		
	claim	**court**	**course**	**dish**
claim	38	0	0	0
court	1	0	0	0
course	1	0	0	0
dish	0	0	0	0

partisjon 10		Antall korrekte: 35/36 Prosentvis presisjon: 97,2 %		
	claim	**court**	**course**	**dish**
claim	35	1	0	0
court	0	0	0	0
course	0	0	0	0
dish	0	0	0	0